Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

Alan da Silva Romualdo

Multimodal classification for detecting products that do not comply with the Americanas S.A.'s marketplace sales policies

São Carlos
2022

Alan da Silva Romualdo

Multimodal classification for detecting products that do not comply with the Americanas S.A.'s marketplace sales policies

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Metodologias e Técnicas de Computação

Supervisor: Profa. Dra. Helena de Medeiros Caseli

São Carlos
2022

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Defesa de Dissertação de Mestrado do candidato Alan da Silva Romualdo, realizada em 19/04/2022.

Comissão Julgadora:

Profa. Dra. Helena de Medeiros Caseli (UFSCar)

Prof. Dr. Aline Marins Paes Carvalho (UFF)

Prof. Dr. Aldebaro Barreto da Rocha Klautau Junior (UFPA)

O Relatório de Defesa assinado pelos membros da Comissão Julgadora encontra-se arquivado junto ao Programa de Pós-Graduação em Ciência da Computação.

Este trabalho é dedicado aos meus amigos que estão sempre presentes me apoiando e a entusiastas de aprendizado de máquina e da computação em geral.

Acknowledgements

Agradecimentos ao Departamento de Computação da UFSCar e à empresa parceira Americanas S.A., a qual incentiva e fomenta essa e outras pesquisas na UFSCar. Em especial à parceria no projeto de extensão "Dos dados ao conhecimento: extração e representação de informação no domínio do e-commerce" (Projeto de extensão #23112.000186/2020-97), do qual este trabalho faz parte.

À minha orientadora Profa. Dra. Helena de Medeiros Caseli, pela sua orientação cirúrgica e pelos seus ensinamentos muito valiosos que me trilhou a ter uma evolução gritante em minha formação científica e pessoal.

Aos meus amigos, João Pedro, Marisa, Biabiriba, Gusta e outros amigos que sempre me fizeram companhia no Discord superando a situação da pandemia, me apoiando e fazendo parceria no meu dia à dia.

Aos meus pais que amo muito e que me possibilitaram trilhar essa caminhada me dando todo suporte que um filho poderia ter.

À minha vó Alice que carrego comigo no coração por todas minhas conquistas e minha tia Eliane, sendo uma segunda mãe para mim.

Às minhas irmãs. Que sempre me contagiam com suas alegrias e amo muito!

Aos amigos da UFSCAR e Americanas S.A. que me repassam muito conhecimento em todas as nossas reuniões ou conversas casuais.

Enfim, às pessoas incríveis que conheci durante esse período de metrado (que foi no período da pandemia) e que mesmo assim, não deixaram de manter o coração quentinho e me influenciar de alguma forma positiva para esse trabalho.

Muito obrigado!

"O importante é não parar de questionar.
A curiosidade tem a sua própria razão para existir"
(Albert Einstein)

Resumo

O aprendizado multimodal para o domínio do *e-commerce* possui alguns métodos de classificação para categorização, recuperação de informação e recomendações de produtos, que geralmente são compostos por modalidades diferentes: imagens e textos. Devido a grandes variações das características dessas modalidades nos produtos, a ausência ou incompletude de informações (por exemplo, atributos de produtos incompletos), há uma dificuldade de métodos de classificação de trabalhar essas informações para melhorar sua classificação, utilizando então a fusão dos recursos em nível de decisão para classificadores específicos de cada modalidade. Assim, este trabalho foi realizado com o objetivo de investigar o aprendizado multimodal entre modalidades visual e textual para o domínio do *e-commerce*. Os experimentos descritos neste documento apresentaram bons resultados para classificação de produtos das categorias "Adulto" e "Dispositivos Ilegais", que fazem parte do conjunto de dados fornecido pela empresa parceira deste projeto. Nesses experimentos, realizou-se o treinamento para as modalidades específicas, com modelos para texto e para imagem, bem como a fusão das duas modalidades em um modelo multimodal. Os melhores modelos foram os textuais de classificação binária produzidos levando em consideração títulos e descrições de produtos: `TD bin-adult` (com recall de 98%) e `TD bin-illegal` (com um recall de 95%). Temos alguns insights sobre a classificação multimodal das classes principalmente para a modalidade visual que, devido à sua natureza, não pode capturar padrões tão bem quanto os modelos textuais.

Palavras-chave: *e-commerce*, classificação multimodal, texto, imagem, produtos.

Abstract

Multimodal learning for the e-commerce domain, some classification methods are needed for categorization, information retrieval and product recommendations, which are generally composed of different modalities: images and texts. Due to large diversification in the characteristics of these modalities or the absence/incompleteness of information (for example, incomplete product attributes), classification methods face many difficulties in dealing with this information in order to improve their classification. Thus, this work was carried out to investigate the multimodal learning in visual and textual modalities for e-commerce. Our experiments show good results for classification of products from "Adult" and "Illegal Devices" categories, which is part of the dataset provided by the partner company of this project. In these experiments, training was carried out for the specific modalities, deriving text and image models, as well as the fusion of the two modalities in a multimodal model. The best models were the binary textual models trained taking into account product titles and descriptions: `TD bin-adult` (with a recall of 98%) and `TD bin-illegal` (with a recall of 95 %). We have some insights about the multimodal classification, mainly for the visual modality which, regarding its nature, could not capture patterns as well as textual models.

Keywords: e-commerce, multimodal classification, text, image, products.

List of Figures

List of Tables

Glossary

ANN Artificial Neural Network

ANNs Artificial Neural Networks

BERT *Bidirectional Encoder Representations from Transformers*

CNN *Convolutional Neural Network*

CCA Canonical Correlation Analysis

CV Computational Vision

DCCA Deep Canonical Correlation Analysis

DBM *Deep Boltzmann Machines*

DCL *Destruction and Construction Learning*

FN False Negative

FP False Positive

HTML *HyperText Markup Language*

KCCA Kernel Canonical Correlation Analysis

LSTM Long short-term memory

LSI Latent Semantic Indexing

ML Machine Learning

NLP Natural Language Processing

RBM *Restricted Boltzmann Machines*

RNN Recurrent Neural Network

RNNs Recurrent Neural Networks

TP True Positive

TN True Negative

Contents

Chapter 1

Introduction

According to the 44th edition of Ebit Nielsen's Webshoppers report[1], e-commerce in Brazil reached new highs, with sales of R$53.4 billion in the first half of 2021. And the importance of e-commerce for businesses and the general public has grown significantly, particularly during the isolation caused by the pandemic situation.

One reason for this growth is the marketplaces which became a crucial alternative for many stores, including supermarkets, to maintain their income during the isolation in the pandemic situation (between 2020-2021). According to the Brazilian Electronic Commerce Association[2], marketplaces accounted for 78% of total e-commerce revenue in 2020. According to the most recent Beyond Borders EBANX[3] report, e-commerce in Latin America grew by 41% in 2021 and is expected to grow by 30% per year until 2025.

Marketplaces often allow sellers to automatically enter any product into the sales platform. This adaptability is required to cover the wide range and quantity of products advertised. However, it can lead to compliance issues if a seller enters a product that does not comply with the marketplace's sales policies. Some products may be sold in certain countries but not in others; some sellers may be authorized to sell or resell specific branded products while others are not; and each marketplace has its own sales policies that may prohibit the sale of certain items. As a result, each marketplace has distinct compliance requirements to ensure a fair assortment.

Companies use a variety of methods to ensure that their assortment is as accurate as possible, including process-based business rules, manual product inspection, and automatic product detection. This study looks into the case of Americanas S.A., Latin America's

[1] <https://www.ecommercebrasil.com.br/noticias/e-commerce-no-brasil-bate-> <recorde-e-atinge-r-53-bilhoes-ebit-> <nielsen-webshoppers/>

[2] <https://abcomm.org/noticias/marketplaces-crescimento-exponencial-ao-> <longo-da-pandemia/>

[3] <https://business.ebanx.com/en/resources/beyond-borders-2021-2022>

largest marketplace with over 200 million active offers. Taking into account the scenario's complexity and breadth, this work investigates machine learning approaches that can be used to accurately detect products in disagreement with Americanas S.A.'s sales policies.

According to Baltrušaitis, Ahuja e Morency (2019), in Machine Learning (ML), a computer system can "learn" by means of algorithms with the ability to catch some patterns based on pattern recognition approach. These ML algorithms have the ability to learn to reproduce an action (or represent knowledge) from examples and, as an output, generate a computational model. Such algorithms can be applied to areas such as Natural Language Processing (NLP), which investigates, proposes and develops different methods, resources and tools capable of dealing with content in natural language; and Computational Vision (CV), which has a similar purpose for images and videos.

Although these ML algorithms are typically used in NLP and CV applications separately, multimodal processing methods, that is, methods capable of dealing with multiple modalities at the same time, were proposed in recent years. For (BALTRUsAITIS; AHUJA; MORENCY, 2019), a modality can refer to how something happens or is experienced, such as NLP, which can be written or spoken; visual signals, which are frequently represented by images or videos; and sound signals, which encode information sounds, vocal expressions, and prosody, including intonations and speech rhythms.

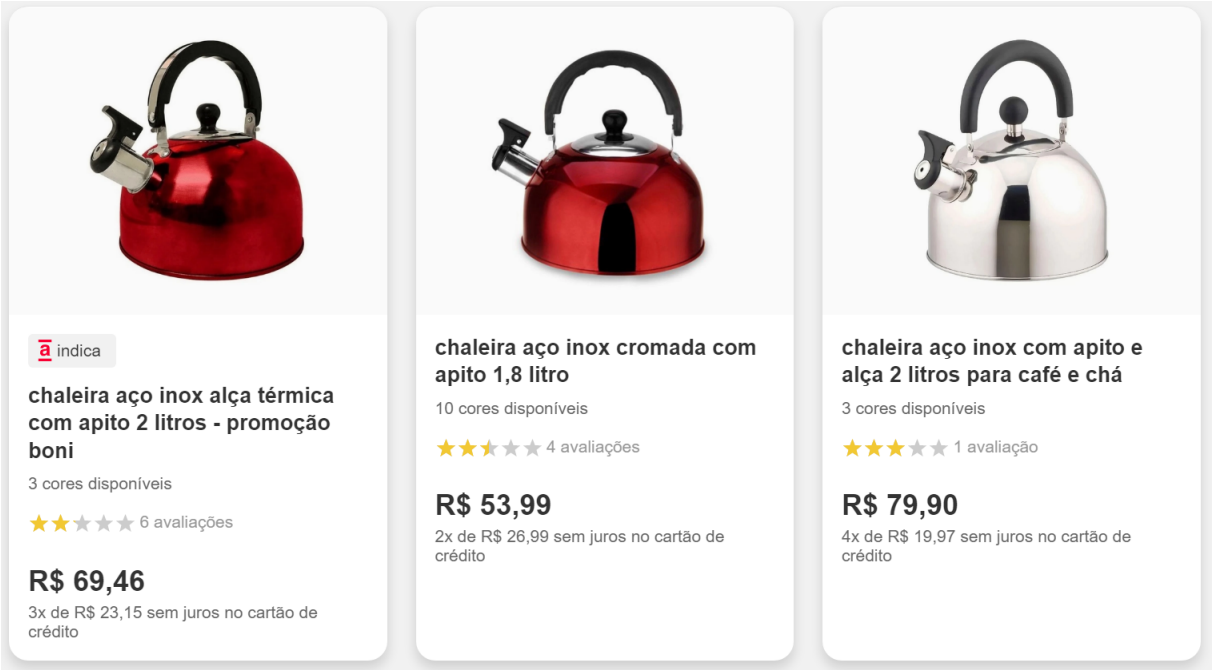
In this scenario, **multimodal machine learning** emerges, with the goal of developing perceptual models for more than one modality, that is, models that are capable of processing and relating information across modalities. According to Wu et al. (2020), large amounts of multimodal data – such as those present in marketplaces, social networks, video platforms, and so on – are currently available on the Internet or in everyday life. As Formal et al. (2020) point out, many techniques have been developed to explore visual information and improve the quality of multimodal application results (visual-textual).

In the specific domain of marketplace, multimodal tasks focus on: **large-scale classification** (text and image), where the goal is to predict the correct category of some product; information retrieval, where the objective, for example, is to retrieve more relevant images or titles and descriptions for a given product; or recommendation systems, in which one can compare different information such as product specifications, purchase options, or even perform visual-textual similarity scoring between those products using various techniques of NLP or CV (AMOUALIAN et al., 2021).

The majority of the information in this domain (marketplace) is conveyed in the form of natural language text or descriptive images and videos of the products for sale. One of the challenges for the marketplace's multimodal tools in dealing with this information is in the seller's customized product advertisements, in which they are allowed to use their own images or product descriptions for their advertisements. Because of these peculiarities, information is presented in an unstructured and dispersed manner, making it difficult to group or classify information from similar products.

To automatically organize, group, or classify these products, techniques that can interpret information in different modalities must be used. Figure 1 shows an example of this problem.[4] When attempting to automatically group these products, the variety in their presentation makes it difficult to detect that they are similar products, since the images have different colors and sizes and were obtained from different angles. Furthermore, a variety of terms and characteristics are mentioned in the textual descriptions.

Figure 1 – Kettle with whistle



Source: Image taken from the Americanas website.

Still using Figure 1 as an example, a binary classification algorithm would classify these products as similar based on the characteristics recognized in the training[5]. This is significant because not all products are necessarily grouped by the information and specific characteristics contained in the modalities, but rather by a multimodal semantic relations between these products.

For example, in the "Adult" category of the dataset used in this work (described in Chapter 6), products are not always linked to specific terms or visual characteristics, and even when there are different intrinsic characteristics like colors and dimensions, there are still numerous variations of products in this category. Examples of products found in the Adult class are: books, sexual objects, content related to violence, or anything that has been classified as adult due to the idea of being adequate only for people older than 18. In addition to the Adult class, this work also considered the Illegal Devices one, which contains items that do not comply with the partner company's sales policy due to the

[4] The images of kettles and other information were obtained from the Americanas website in March 2022: <https://www.americanas.com.br/>.

[5] According to (BALTRUsAITIS; AHUJA; MORENCY, 2019), a computer system can understand a modality by means of the use of algorithms that can learn through pattern recognition.

purpose of the product in spite of their textual or visual features. Examples of items in this class are signal blockers, which can be for GPS, radars, or cell phone signals; products with spying intentions, such as hidden cameras in pens, lamps, decorative objects, among others. Although, in this work, both classes (Adult and Illegal Devices) can be joined in just one of non-compliant products, we decided to maintain this original division from Americanas S.A. to be able to perform multiclass experiments.

The approaches in the domain of e-commerce, described in the Chapter 3 and cited in (AMOUALIAN et al., 2021), challenge themselves in the task of classification in an irregular data distribution. To deal with this problem, efforts have been made in two different ways: (1) unimodal approaches and (2) multimodal approaches (BI; WANG; FAN, 2020). Multimodal approaches combine modalities by fusion techniques for multimodal learning and can be grouped by resource or decision-level fusion (ZAHAVY et al., 2018).

Bi, Wang e Fan (2020) use both strategies, with the decision-level merger outperforming the resource-level one. Furthermore, the authors used distinct classifiers for each text and image modality that are combined in the *late fusion* strategy, that combines features from different modalities at or closer to the classification layer. Chordia e Kumar (2020) use models for specific modality, but they also use a *Co-attention* technique proposed by (LU et al., 2016), which they attribute an important contribution to the overall performance of their proposal. CamenBERT[6], a model based on the Roberta architecture (LIU et al., 2019) trained for French, was used in these two works.

The *late fusion* approach was also used by Wirojwatanakul e Wangperawong (2019) to categorize Amazon's e-commerce products. They trained separate models for each modality (text, image, and description), which they named the tri-modal fusion model. The authors created an architecture with three components in Zahavy et al. (2018): (1) an CNN of (KIM, 2014) for the text, (2) an CNN of (SIMONYAN; ZISSERMAN, 2015) for image, and (3) a decision neural network, which learns to choose classification between these two modalities. In general, all models use CNNs for images, such as VGG (SIMONYAN; ZISSERMAN, 2015), ResNet (HE et al., 2016), or CNN from (KIM, 2014), as well as some NLP methods, such as some architecture by *Bidirectional Encoder Representations from Transformers* (BERT) (DEVLIN et al., 2018) or *embeddings* generated by commonly used methods, such as Glove (PENNINGTON; SOCHER; MANNING, 2014).

In the experiments presented in Chapter 6, the two fusion approaches were used: the resource level one (*early-fusion*) for title and description; and the *late-fusion* for combining the resulting text with the image template.

[6] url<https://camembert-model.fr>

1.1 Objective

The goal of this work is to investigate the unimodal and multimodal classification of products based on their textual and visual information. The classification task investigated here is that of distinguishing between binary negative class products that comply and positive classes (adult and illegal) products class that do not comply with the Americanas S.A.'s sales policies. This work has the hypothesis that multimodal models would perform better than unimodal ones in this task. To test our hypothesis we experimented with products from Adult and Illegal Devices classes in a dataset provided by Americanas S.A. We also investigated binary, multiclass and ensemble approaches to deliver the classification models.

For this purpose, product datasets containing textual (titles and descriptions) and visual (photos and images) information were used. As already mentioned, this data was provided by Americanas S.A., which is a partner company in this project. Americanas S.A. is a digital company that owns the pioneering brands in the context of Brazilian e-commerce: Americanas[7], Submarino[8], Shoptime[9] and Sou Barato[10].

1.2 Text organization

This Master Thesis is organized as follows. Chapter 2 presents: the problem of multimodal classification and its challenges; the unimodal and multimodal representations usually applied for text and image; and the main automatic evaluation measures used in classification tasks. Chapter 3 presents the methods proposed in the literature selected as the most relevant to deal with the problem of multimodal classification in e-commerce. Chapter 4 provides a more detailed description of the problem, the materials used and the methodology followed to solve the problem of multimodal classification of non-compliant products. The error analysis tool developed in this work to add the qualitative evaluation is presented in Chapter 5. The results of all the experiments carried out with uni and multimodal models are presented in Chapter 6. Finally, Chapter 7 presents the conclusions of this work pointing out the main contributions and some possible future works.

[7] <https://www.americanas.com.br/>
[8] <https://www.submarino.com.br/>
[9] <https://www.shoptime.com.br/>
[10] <https://www.soubarato.com.br/>

Chapter 2

Multimodal Learning

For the multimodal classification in the *e-commerce* domain to occur, we believe that it is important that the multimodal learning methods be able to absorb the characteristics present in each modality (visual and textual) in product.

To illustrate this problem, consider the two products for sale on the Americanas website[1] shown in Figure 2 and additional information about them in Figures 3 and 4, respectively.

A human would be able to detect contextual similarities between products based on multimodal analysis of text and image, but the question we raise here is: would automatic processing of each modality separately (or even together), reach the same conclusion?

The challenges present in this multimodal learning (or representation) process are the subject of Section 2.1. Section 2.2 deals specifically with unimodal and multimodal representations for text and images. Finally, in Section 2.3, measures frequently used in the automatic evaluation of multimodal information classification and retrieval systems are explained.

2.1 Challenges of multimodal learning

According to Baltrušaitis, Ahuja e Morency (2019), multimodal learning has five main challenges that came from the heterogeneity of data in the different modalities of text, speech or vision for the understanding of natural phenomena. These challenges are named as: *Representation*, *Translation*, *Alignment*, *Fusion* and *Co-learning*.

[1] The images of kettles and other information present in the Figures 1, 2, 3 and 4 were taken from the Americanas website (<https://www.americanas.com.br/>) in June 2020.

Figure 2 – Example of products that should be considered similar.

Source: Images taken from the Americanas's website.

Figure 3 – Additional information about the first (on the left) kettle.

informações do produto

Qualidade e elegância!

Com os produtos da Eirilar ficou mais prático, as chaleiras em inox dão um toque a mais de estilo para a sua cozinha, além de serem feitas em aço inox.

Descrição:
Uma moderna e prática chaleira para fazer seu chazinho da tarde! Totalmente em inox, vem com cabo ergonômico superior que impede que o calor aqueça e você queime as mãos. Ferva água com estilo e elegância e renove sua cozinha!

Prepare e sirva chá de acordo com o seu gosto... mais qualidade e praticidade à você!

Especificações:

(C x L x A): 20 x 20 x 20 cm. / 2,5 litros

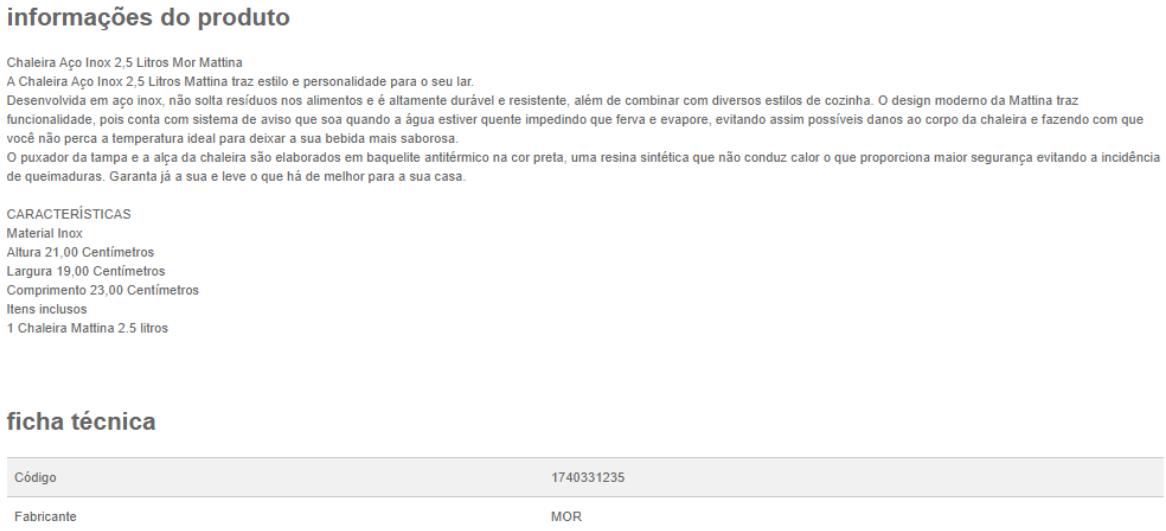
ficha técnica

Código	8575540
Código de barras	7896747150213
Cor	Vermelho
Capacidade (L)	2,50
Fabricante	Eirilar
Composição/Material	Inox

Source: Information taken from the Americanas's website.

The *Representation* challenge is defined as being the most fundamental to multimodal machine learning. In this challenge, it is necessary to know how to represent and summarize multimodal data in order to explore the **complementarity** and **redundancy** of multiple modalities. For example, while product descriptions use phrases and words, the

Figure 4 – Additional information about the second (on the right) kettle.



informações do produto

Chaleira Aço Inox 2,5 Litros Mor Mattina
A Chaleira Aço Inox 2,5 Litros Mattina traz estilo e personalidade para o seu lar.
Desenvolvida em aço inox, não solta resíduos nos alimentos e é altamente durável e resistente, além de combinar com diversos estilos de cozinha. O design moderno da Mattina traz funcionalidade, pois conta com sistema de aviso que soa quando a água estiver quente impedindo que ferva e evapore, evitando assim possíveis danos ao corpo da chaleira e fazendo com que você não perca a temperatura ideal para deixar a sua bebida mais saborosa.
O puxador da tampa e a alça da chaleira são elaborados em baquelite antitérmico na cor preta, uma resina sintética que não conduz calor o que proporciona maior segurança evitando a incidência de queimaduras. Garanta já a sua e leve o que há de melhor para a sua casa.

CARACTERÍSTICAS
Material Inox
Altura 21,00 Centímetros
Largura 19,00 Centímetros
Comprimento 23,00 Centímetros
Itens inclusos
1 Chaleira Mattina 2.5 litros

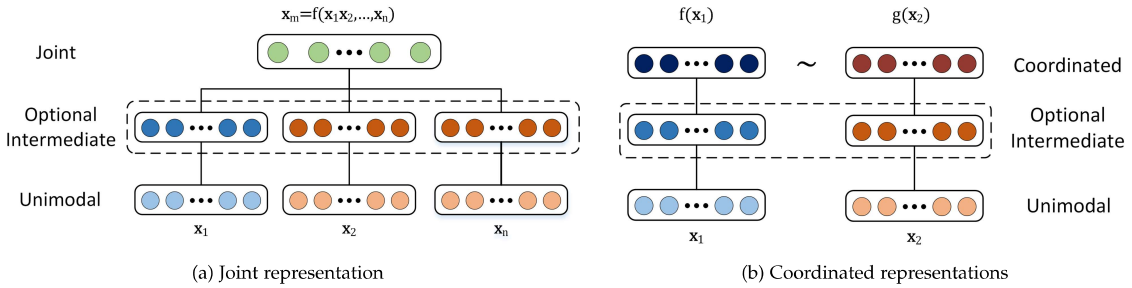
ficha técnica

Código	1740331235
Fabricante	MOR

Source: Information taken from the Americanas's website.

visual modality is commonly represented as signs, videos or images. This fact causes data heterogeneity and makes it difficult to build such representations. To deal with the representation challenge there are techniques that map multimodal data to: (i) the same space (joint space or *joint embedding*), or (ii) the coordinate space (*coordinated embedding*). In the joint representation illustrated in Figure 5 (a), unimodal signals are combined in the same representation space, while in the coordinated one, illustrated in Figure 5 (b), unimodal signals are processed separately but with the reinforcement of some similarity constraints to bring them to a coordinated space.

Figure 5 – Joint representation and coordinated representation.



(a) Joint representation (b) Coordinated representations

Source: (BALTRUsAITIS; AHUJA; MORENCY, 2019).

The *Translation* challenge is related to determining how to map or convert data from one modality to another. This translation encompasses not only the heterogeneous untranslated data, but also the correlation between modalities, which is usually open or subjective. For example, there are several correct ways to describe a product's image, and these descriptions by themselves may not be enough to correlate two products. In the text there may be a specification (such as size) that is not present, or not easily

detected visually in the image.

The *Alignment* challenge seeks to identify the direct mappings between (sub) elements of two or more modalities. For example, how to identify which part of the text is the description of product's color and associate (align) that part of the textual modality to the corresponding part in the visual modality?

The *Fusion* is the challenge of joining information from two or more modalities to make a prediction. For example, in an image of a red product this characteristic may not be present in the textual description of that same product. In this case, the fusion is like a transition from visual modality to natural language in order to fill in missing information in the textual description of the product. It is a prediction considering the different noises or missing information that may affect this joining.

Finally, the *Co-learning* challenge explores how knowledge learning about one modality can help a computational model trained based on a different modality. According to Baltrušaitis, Ahuja e Morency (2019), this challenge is especially relevant when there is a lack of resources in one of these modalities. Unlike merging modalities, co-learning occurs in a way that uses one modality to train models capable of understanding another modality. Initially, it was more interesting for us to verify the unimodal representation combined with the fusion technique.

2.2 Representation of texts and images

Multimodal learning emerged from the premise that unimodal representation alone is not enough to model a complete set of human concepts and perceptions. Therefore, the multimodal representation becomes necessary for the representation of the concepts present in multimodal data. For example, the concept of "Stylish and beautiful product" may be based on vision and, as a consequence, may be difficult to describe using only natural language.

To understand representations involving multiple modalities, the next subsections describe how unimodal and multimodal representations are generated from texts and images.

2.2.1 Unimodal representations

In textual representations, a vocabulary is usually built on the basis of some constraints – such as occurrence frequency – applied to select words and even expressions , depending on the application. Then, this vocabulary is used for two types of representation approaches: distributive or distributed.

In the distributive approach, illustrated in Figure 6, a sentence is represented, for example, by a vector that indicates the occurrence frequency of the words that are present in the sentence, disregarding the relation and order of these words. In the distributed

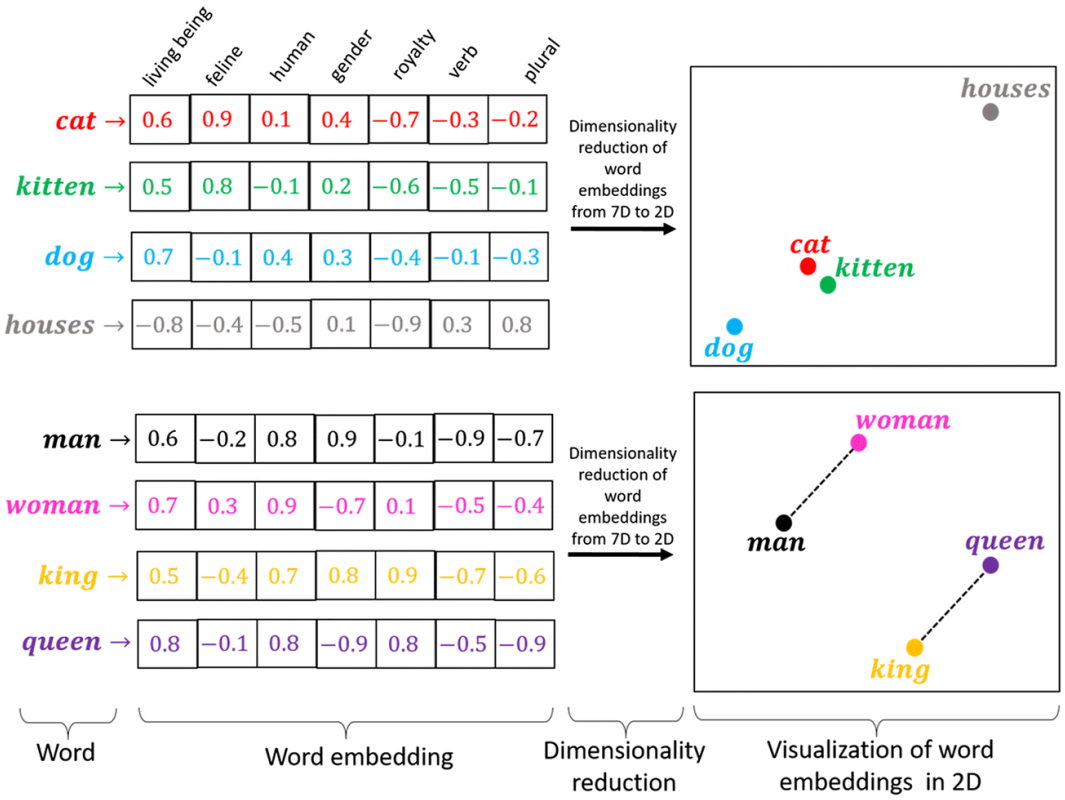
approach, illustrated in Figure 7, the representation is more sophisticated as it may capture syntactic and semantic relations for each word in the vocabulary. According to (LOCHTER, 2021), the main difference between these two types of representation lies in the ability of the distributed ones (*word embeddings*) to catch information based on the context of co-occurring words, which usually leads to a closer proximity between semantically similar words because they tend to occur in the same context.

Figure 6 – Example of textual representation with *bag-of-words* (distributive approach).

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

Source: (ZHOU, 2019).

Figure 7 – Example of fictional textual representation with *word embeddings* (distributed approach).

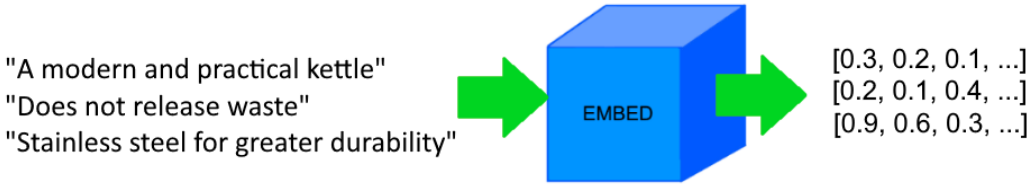


Source: (GAUTAM, 2020).

Examples of methods used in textual representation are the traditional Latent Semantic Indexing (LSI), the ones widely used today such as Word2Vec (MIKOLOV et al., 2013), FastText (BOJANOWSKI et al., 2017) and Glove (PENNINGTON; SOCHER; MANNING, 2014), and those considered the *state of art*, which are contextualized lan-

guage models like BERT (DEVLIN et al., 2018). As an illustration of a possible vector representation generated by these methods, in Figure 8 we can see the representations for some of the sentences of Figures 2, 3 and 4.

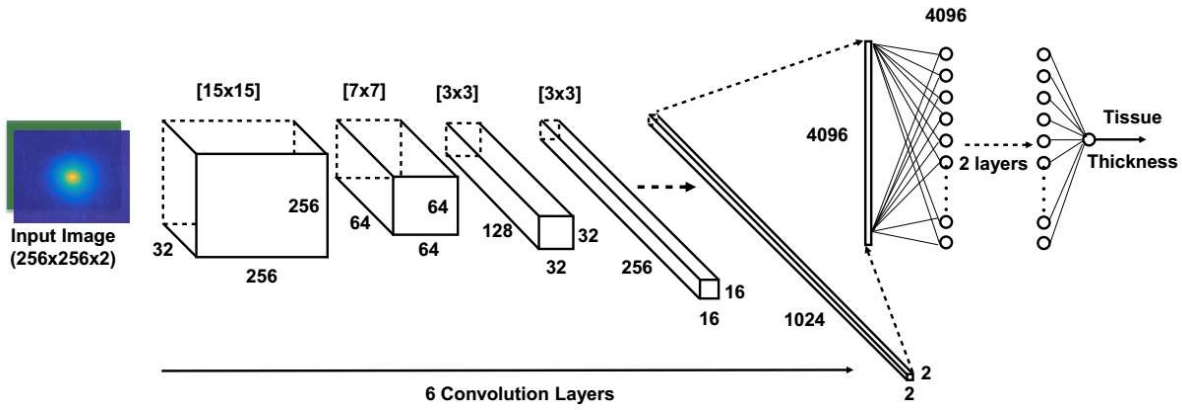
Figure 8 – Example of a fictional unimodal representation of text.



Source: The author.

For visual representation, Artificial Neural Networks (ANNs) have been widely used to generate an image representation as output values of final layers of CNN. The CNN (Figure 9) is a type of Artificial Neural Network (ANN) used in many applications of Digital Image Processing. According to Zhang et al. (2020), there is greater semantics when combining convolutional features with labels for regions of objects found in the image. Thus, to generate the image representation, some popular networks can be used such as AlexNet (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), VGG Networks (SIMONYAN; ZISSERMAN, 2015), GoogLeNet (SZEGEDY et al., 2015), ResNet (HE et al., 2016) and also R-CNN (GIRSHICK et al., 2014), which combine convolutional features with labels for regions of objects found in the image.

Figure 9 – Example of CNN with fully connected layers.



Source: (MANIT; SCHWEIKARD; ERNST, 2017).

2.2.2 Multimodal representations

Based on the example of Figure 2, one question that arises is: "how to represent and summarize multimodal product data in order to explore the complementarity and

redundancy of the information present in its various modalities?". The most common approaches are to map to the same space (joint approach) or coordinate modalities (coordinate approach), making it possible to retrieve information based on data from different modalities.

In the joint approach, ANN can be used for multimodal learning of visual and textual data. For that, the ANNs are trained to recognize the products or objects in the images or texts, where given the multilayer nature of a deep neural network, each successive layer is hypothesized to represent the data in a more abstract way (see Figure 9). Consequently, it is possible to use the last neural layer (or the one before it) as a form of representation.

To build a multimodal representation using ANNs, each modality starts separately with several individual neural layers followed by a hidden layer that projects the modalities into a joint space. The joint multimodal representation is then passed to several hidden layers or used directly for prediction. Such models can be trained end-to-end, learning at the same time how to represent the data and to perform a specific task. In this way, when neural networks are used, we have a close relation between multimodal representation learning and multimodal fusion.

Probabilistic models can also be used to build representations by using latent random variables to represent data at a higher level of abstraction, using models such as *Deep Boltzmann Machines* (DBM) and *Restricted Boltzmann Machines* (RBM). One of the advantages of using multimodal DBM for learning multimodal representations is its generative nature, which provides an easy way to deal with missing data. Their biggest disadvantage is the difficulty to train them (BALTRUsAITIS; AHUJA; MORENCY, 2019).

Sequential models of Recurrent Neural Networks (RNNs) like Long short-term memory (LSTM) are commonly used to represent variable-length sequences, such as sentences, videos, and audio streams. According to Kiros, Salakhutdinov e Zemel (2014), LSTM is one type of Recurrent Neural Network (RNN) that has special units, called memory blocks, in a recurrent hidden layer that are used to store information and explore the long-range context. These memory blocks have self-connecting memory cells that store the temporal state of the network, in addition to special multiplicative units, called gates, used to control the flow of information. They are also known to be surrounded by locking units that serve to read, write and reset information. LSTMs networks have been used to achieve cutting-edge performance in various tasks, such as handwriting recognition, sequence generation, speech recognition and machine translation, among others.

In the coordinated approach, the products can be represented through the structure constraint, where restrictions of similarities between the modalities are imposed so that there is a representation of the same object in the different modalities used. For example the *cross-modal hashing-compression* method represents the data in compact binary codes so them to be similar binary codes, for the same objects, across different modalities. In

this coordinated representation approach, three methods are presented: Canonical Correlation Analysis (CCA), Kernel Canonical Correlation Analysis (KCCA) and Deep Canonical Correlation Analysis (DCCA). These methods are unsupervised and only optimize correlation over representations, capturing mainly what is shared between modalities.

After training and representation of the modalities by these methods, alignment occurs. According to Baltrušaitis, Ahuja e Morency (2019), alignment can be made in two different ways:

- **Explicit alignment**: In this kind of alignment, there is an explicit interest in aligning the sub-components of the modalities, as in the alignment of the textual characteristics described in the products' descriptions with the regions or objects detected in their images. A very important part of explicit alignment is the calculation of similarity between modalities and how it is done. To do so we have unsupervised methods, in which training is performed assuming some constraints (such as the temporal order of the sequence or the existence of a measure of similarity between the modalities); or supervised ones, in which a measure of similarity is trained from annotated data and the model learned from this measurement is later used for alignment. In this type of alignment, classification models could receive textual features and regions of interest or objects detected in the images to detect non-compliant products.

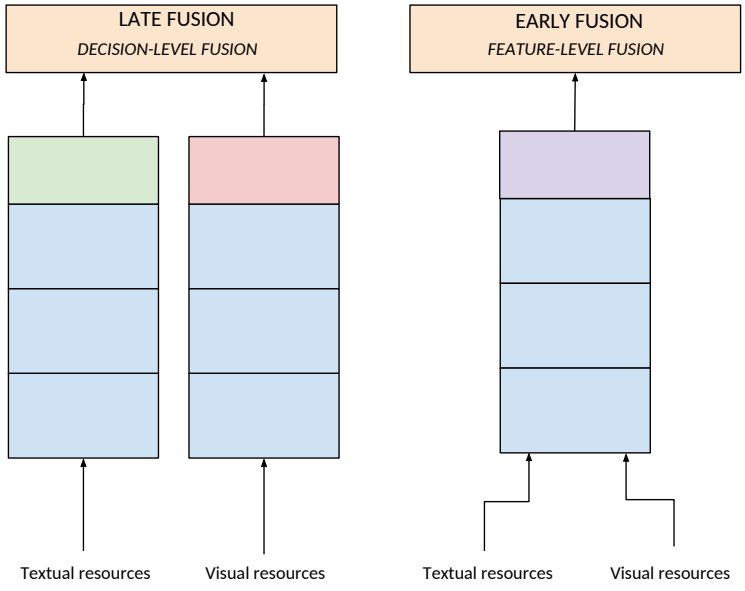
- **Implicit alignment**: This kind of alignment occurs as an intermediate (usually latent) step of another task, for example, in retrieving a product image from a textual description there may be an alignment step between the words of the description and the regions of image. These models do not explicitly align the data, nor are they based on alignment examples, but learn how to latently align the data during model training. A very popular way to perform the alignment implicitly is via attention mechanisms, which allow the decoder to focus on the sub-components of the source instance which are most relevant to the task at hand. Again, it could be used in multimodal classification as a mechanism that deals with filling missing information.

Fusion is the essential research topic in multimodal studies. It seeks to unify information from unimodal data from two or more modalities into a single compact multimodal representation (ZHANG et al., 2020). According to Zhang et al. (2020), an approach falls into the fusion category if it focuses on architectures to integrate unimodal representations for a specific task. Thus, these fusion methods can be divided based on the moment at which the fusion occurs. There may be fusion at the beginning and at the end of the processing of modalities. In Figure 10 it is possible to see the two most common types of fusion. Recent studies have focused on intermediate methods that allow fusion to take

place in multiple layers of a deep model. In the work of Zhang et al. (2020), the fusion approaches are divided into:

- **Simple Operation-base Fusion**: In this approach, vector features from different modalities can be integrated using simple operations such as concatenation or weighted sums, which often have few or no associated parameters.
- **Attention-based Fusion**: This approach refers to methods of weighted sum of a set of vectors with scalar weights that are dynamically generated by an attention model associated with each part/step of the processing.
- **Bilinear Pooling-based Fusion**: This approach is frequently used for merging vectors of visual features with vectors of textual features in order to create a joint representation space by computing their outer product, which facilitates multiplicative interactions between all elements in the two vectors. Bilinear clustering generates a two-dimensional representation by linearizing the matrix generated by the outer product into a vector, which means that this method has more ability to discriminate between visual and textual features.

Figure 10 – The two types of fusion as cited by (BI; WANG; FAN, 2020).



Source: The author.

2.3 Automatic evaluation

This section presents some of the main evaluation measures used in the literature to evaluate classification methods. All of them will be explained based on the Table 1, in

which the relation between the *expected class* and the *predicted class* guides the evaluation of the items.

Table 1 – Confusion matrix

		Classification predicted by the model	
		True	False
Expected prediction	True	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

To illustrate the calculation of each measure, we will use the example from Table 2. That table simulates values of classification for a fictitious dataset with 90 instances, being divided equally into two classes "true" and "false" instances. The fictitious model ranked 55 instances as "True" and 35 as "False". Out of 45 values expected as "True", only 35 were correct (TP) and that 20 that were expected as "False " were classified as "True" (FP). Also, another 10 that should be "True" were incorrectly classified by the model as "False" (FN) against 25 (TN) that were correctly classified as "False".

Table 2 – An example of a confusion matrix for fictitious classification result.

		Classification predicted by the model	
		True	False
Expected prediction	True	35 instances	10 instances
	False	20 instances	25 instances

2.3.1 Accuracy

The objective of accuracy is to indicate the performance of the model by dividing all classifications performed correctly over all classifications, thus having the following formula:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Using Table 2 as an example, the accuracy value is:

$$Acc = \frac{35 + 25}{35 + 25 + 20 + 10} = \frac{60}{90} = 67\%$$

2.3.2 Precision

Precision seeks to inform that among all the positive hits that the model makes (predicted classification), how many are correct (TP). Its formula is:

$$Precision = \frac{TP}{TP + FP}$$

Using Table 2 as an example, the precision value is:

$$Precision = \frac{35}{35 + 20} = \frac{35}{55} = 64\%$$

2.3.3 Recall

Also called sensitivity, this evaluation measure tells us how many positive instances are correct among all predicted class situations. See the formula:

$$Recall = \frac{TP}{TP + FN}$$

Using Table 2 as an example, the recall value is:

$$Recall = \frac{35}{35 + 10} = \frac{35}{45} = 78\%$$

2.3.4 F1-Score

F1-Score is the harmonic mean between the precision (2.3.2) and the recall (2.3.3) of the model. It has the following formula:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Using Table 2 as an example and the accuracy and recall values obtained before, the F1 value is:

$$F1 = 2 \times \frac{0.64 \times 0.78}{0.64 + 0.78} = \frac{0.49}{1.42} = 34.5\%$$

2.3.4.1 *Micro Averaged*

The *Micro Averaged* or *Micro F1-Score* is the *F1-Score* adapted to evaluate the performance of models in different data sets, when dataset sizes are variable. Having the following formulas:

$$MicroPrecision = \frac{TP_1 + TP_2}{TP_1 + FP_1 + TP_2 + FP_2}$$

$$MicroRecall = \frac{TP_1 + TP_2}{TP_1 + FN_1 + TP_2 + FN_2}$$

$$MicroAverage = 2 \times \frac{MicroPrecision \times MicroRecall}{MicroPrecision + MicroRecall}$$

Again using Table 2 as an example for classification results 1, consider the following values for a second set, the 2:

$$TP_2 = 37 \quad TN_2 = 50 \quad FP_2 = 30 \quad FN_2 = 43$$

So we have:

$$MicroPrecision = \frac{35 + 37}{35 + 20 + 37 + 30} = \frac{72}{122} = 59\%$$

$$MicroRecall = \frac{35 + 37}{35 + 10 + 37 + 43} = \frac{72}{125} = 57\%$$

$$MicroAverage = 2 \times \frac{0.59 \times 0.57}{0.59 + 0.57} = 2 \times \frac{0.34}{1.16} = 29\%$$

2.3.4.2 *Macro Averaged*

The *Macro F1-Score* or *Macro Averaged* is also the *F1-Score* (2.3.4) adapted to evaluate the performance of models in different data sets. It has the following formulas:

$$MacroPrecision = \frac{Precision_1 + Precision_2}{2}$$

$$MacroRecall = \frac{Recall_1 + Recall_2}{2}$$

$$MacroAverage = 2 \times \frac{MacroPrecision \times MacroRecall}{MacroPrecision + MacroRecall}$$

Using the previous calculations from Table 2 as an example for results in set 1, consider the following values for set 2:

$$Precision_2 = \frac{37}{37 + 30} = 55\% \quad Recall_2 = \frac{37}{37 + 73} = 34\%$$

$$MacroPrecision = \frac{0.64 + 0.55}{2} = 59\%$$

$$MacroRecall = \frac{0.78 + 0.34}{2} = 56\%$$

$$MacroAverage = 2 \times \frac{0.59 \times 0.56}{0.59 + 0.56} = \frac{0.33}{1.15} = 28\%$$

Chapter 3

Related works

This chapter briefly describes some previous works considered relevant for multimodal learning in the e-commerce domain.

3.1 A Multimodal Late Fusion Model for E-Commerce Product Classification (BI; WANG; FAN, 2020)

Bi, Wang e Fan (2020) investigated a multimodal *late fusion* approach based on text and image modalities to categorize Rakuten's *e-commerce* products[1]. Specific deep neural networks were developed for each input modality (text and image) and different fusion approaches were tested. These authors won first place in the multimodal product classification task of the SIGIR E-Commerce Workshop 2020[2] data challenge.[3]

For the textual representation, the authors used CamemBERT[4], given that most product titles and descriptions are in French. For the image-based product classifier, the authors employed the commonly used network ResNet152 (HE et al., 2015) pre-trained on the ImageNet dataset.

For the data, around 99 thousand products containing the information of: title, description, image(s) and type code (category). In total, there were 27 product categories in the training dataset, ranging from 764 to 10,209 data per category. 84,916 instances were used for training, 937 for the test in a phase 1 and 8,435 for test in a phase 2. Regarding

[1] Rakuten is a Japanese company that offers services such as shopping, online video viewing and more.
[2] <https://sigir.org/sigir2020/>
[3] The code of this work is available at: <https://github.com/Wang-Shuo/SIGIR2020Challenge>.
[4] CamemBERT (<https://camembert-model.fr>) is a state-of-the-art language model for French based on the RoBERTa architecture (LIU et al., 2019) pre-trained on the French sub-corpus of the OSCAR multilingual corpus (<https://oscar-corpus.com>).

data pre-processing strategies, the authors used *cleanlab*[5], a noise reduction tool, to remove the top@10 images from each category that had some label error. According to the authors, these images could disrupt the training as they clearly were incorrectly labeled. For text pre-processing, extra spaces and some *HyperText Markup Language* (HTML) tags were removed from titles and descriptions.

Although there were 27 product categories in the training set, the products belong to only 4 top-level categories (Children, Books, Family and Entertainment). Thus, the authors treated the image-based product classifier as a refined image recognition task and used a method *Destruction and Construction Learning* (DCL)(CHEN et al., 2019)[6].

The authors experimented with two multimodal fusion approaches – early fusion and late fusion – but they focused, in phase 2, on the late fusion approach (shown in Figure 11) since it was the one which had obtained better results in phase 1, as shown in Table 3.

Table 3 – F1 values for unimodal (image and text) and multimodal models in Decision-level (late) Fusion, Feature-level (early) Fusion and the ensemble approaches.

Method	Phase 1	Phase 2
Uni-Image Classifier	69.21	-
Uni-Text Classifier	89.93	-
Feature-level Fusion	89.87	-
Decision-level Fusion	**90.94**	90.17
Ensemble		**91.44**

Source: (BI; WANG; FAN, 2020)

In the decision-level (late) fusion approach (Figure 11), modality-specific classifiers are used as input and then another network is trained from the class probabilities predicted by each modal classifier.

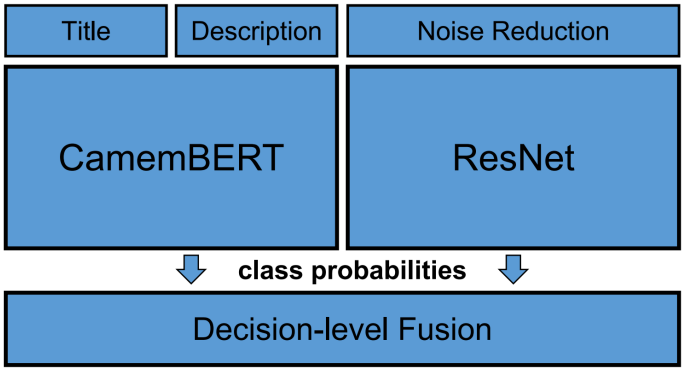
From the results presented in Table 3, it is observed that the decision-level fusion approach achieved better results than the unimodal models and the feature-level fusion approach and that the ensemble strategy achieved an even better result.

In contrast with our work on multiclass and binary classification of non-compliant products, the authors used 26 categories with an irregular frequency of products. Other approaches in textual and visual modalities were used, including CamemBERT, a contextualized language model based on BERT for French, and Resnet 152, which was also pre-trained in ImageNet. The late fusion of this work guided us to follow the same path. The ensemble approach also encouraged our research.

[5] <https://github.com/cgnorthcutt/cleanlab>

[6] According to Chen et al. (2019), experts can distinguish objects only looking at parts of the image. Thus, they proposed this model to increase the *fine-grained* difficulty and push the classification model to acquire specialized knowledge.

Figure 11 – Overview of the multimodal classification model proposed by Bi, Wang e Fan (2020)



Source: (BI; WANG; FAN, 2020)

3.2 Co-Attention Transformer Model (CHORDIA; KUMAR, 2020)

Chordia e Kumar (2020) used language models produced by mono and multilingual Transformers to generate textual representations to experiment with image combinations and a *co-attention* architecture proposed by (LU et al., 2016). Furthermore, the authors also experimented with ensemble that combines different experiments to obtain better F1 values. As the work in the previous subsection (3.1), these authors participated in the SIGIR 2020 E-Commerce Workshop[7], in the task of multimodal classification of products and came in second [8].

They use textual and visual resources extracted from pre-trained models to explore concatenation, bilinear transformation and scalar product techniques. The fused features are transported over a network of two fully connected layers with a *softmax* classifier. Among the different combinations of pre-trained image and text models (Table 4), the 2048-layer model of the ResNet152 network (HE et al., 2015) in combination with CamemBERT had the best performance.

Table 4 – F1 values for pre-trained models.

Image-Model	CamemBERT	FlauBERT	M-BERT
Resnet152	88.78	88.52	87.06
Resnet101_32x8d	88.72	84.30	87.20

Source: (CHORDIA; KUMAR, 2020).

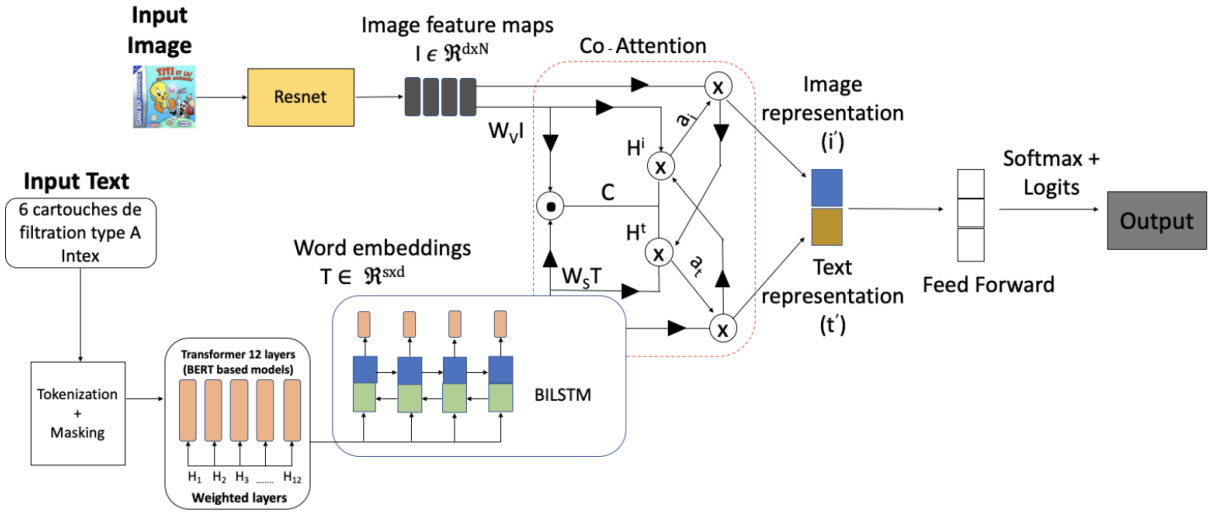
The *Co-Attention Transformer Model* (Figure 12) proposed by the authors has four main components: (1) a pre-trained image model to extract the visual features by means of a ResNet network; (2) a 12-layer BERT-based Transfomer language model that enco-

[7] <https://sigir.org/sigir2020/>

[8] The code for this work is publicly available at: <https://github.com/VarnithChordia/Multimodal_Classification_Co_Attention>

des the text; (3) the Bi Directional Long Short Term Memory (BILSTM), which learns dynamic token dependency sequences from the output of the Transformer and (4) a modification to the co-attention model to act jointly with words and images, and not only words, phrases and sentences as proposed by Lu et al. (2016).

Figure 12 – Model architecture proposed by Chordia e Kumar (2020) which uses co-attention.



Source: (CHORDIA; KUMAR, 2020)

For text pre-processing, alphanumeric tokens, out-of-position symbols and HTML tags were removed. Titles and descriptions were concatenated to create a single corpus of text that served as input to the model. 20% of the training dataset was used as the validation set. From the total of 99,000 products, about 84,000 were in the training set. All products were associated with an image, a title in French and an optional description, like the example in Figure 5.

Table 5 – Example of a training instance (in French) from the product dataset.

Integer_id	Title	Description	Image_id	Product_id
0	Jeep Police - Gevarm-Gevarm	Nan	1193217616	3136702026
1	Court Joyeux No√'l En Peluche Taie ……	No√'l en peluche court Taie Sofa	1323615566	4231863665
2	Sauna infrarouge Largo …..	mensions : 150x105x190 cm ou .	1158121321	2695198357
3	BAGUE POUR LAME SOUS-SOLEUSE G. ET D.	Nan	1096607258	1657064583

Source: (CHORDIA; KUMAR, 2020)

The results of the experiments are presented in Table 6. From these results, the authors of the work concluded that the different modules of the proposed architecture contribute to the performance of the model. For example, switching from BILSTM to a linear layer lowered the F1-score, which indicates that Transformer alone does not increase the score. Also, with the removal of the weighted layers (Figure 12) and considering only the top layer, there was a drop in overall performance which, according to the authors, was due

to the information loss in the lower layers. In short, the architecture proposed by the authors has a better performance than just a direct classification baseline approach as shown in the Table 6.

Table 6 – F1 values obtained when Chordia e Kumar (2020) evaluated the impact of each module on the architecture they proposed.

Model	F1
ResNET152 + CamemBERT	**88.78**
ResNET152 + CamemBERT + W/O BILSTM	76.18
ResNET152 + CamemBERT + W/O weighting layers	83.53
ResNET152 + CamemBERT + text attention alone	80.16
ResNET152 + CamemBERT + image attention alone	86.80

Source: Adapted from (CHORDIA; KUMAR, 2020)

To improve the results, the authors tried an experiment with the combination (ensemble) of the classifiers, as shown in Table 7. According to the authors, the set that obtained the best performance was a combination of all five models.

Table 7 – F1 values for combining multiple models using the co-attention technique.

Model	F1
ReseNET152 + CamemBERT + FlauBERT	90.40
ResNET152 + CamemBERT + M-BERT	90.03
ResNET152 + CamemBERT + M-BERT + FlauBERT	90.95
ResNEXT101 + CamemBERT + FlauBERT	90.38
ResNEXT101 + CamemBERT + M-BERT + FlauBERT	90.88
All 6 models	**91.36**

Source: Adapted from (CHORDIA; KUMAR, 2020)

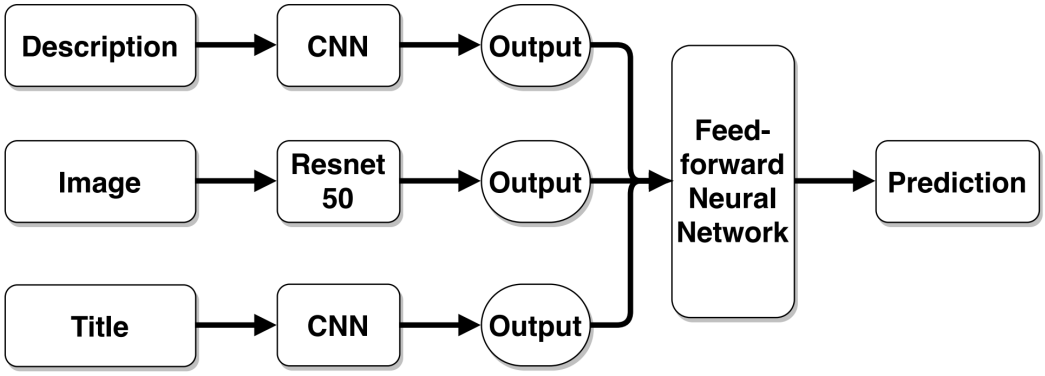
As in our work, the authors also explored simple concatenation for merging the modalities, but explored others as well, such as bilinear transformation and scalar product techniques. The experiments to evaluate the impact of each module on the architecture of this work influenced the development of our error analysis tool (described in Chapter 5), but we look for understanding how our model learned. This work also used co-attention while ours does not, its task was to classify an irregular frequency of products in than 26 categories and this work also addressed ensemble.

3.3 Multi-Label Product Categorization Using Multi-Modal Fusion Models (WIROJWATANAKUL; WANGPERAWONG, 2019)

Wirojwatanakul e Wangperawong (2019) investigated multimodal *late fusion* approaches using images and product descriptions and titles to categorize items on Amazon's

e-commerce. The authors proposed an approach to combine an arbitrary number of modalities, which they called the *tri-modal fusion model* (Figure 13), which considers images, titles and descriptions separately. Also in their research problem, the authors tried to classify products that could be labeled in more than one class and subclasses (multi-label classification).[9]

Figure 13 – Architecture of the *tri-modal* fusion model for product categorization using title, description and image.



Source: (WIROJWATANAKUL; WANGPERAWONG, 2019).

The dataset used by that study has 9.4 million Amazon products (MCAULEY et al., 2015), without the information about the hierarchy of classes or subclasses. From the whole set of products, 119,073 of them were randomly selected and the first 90,000 products were used for training. After pre-processing, a product could belong to up to 122 possible classes, which differentiates this work from the previous ones, due to the possibility of the product having several labels. Each product in the dataset contains the following information: image, description, title, price and co-purchasing network. However, the original paper does not present examples of these data.

During training, the authors used the *Adam optimizer*, *cross-entropy* as loss function and *sigmoid* activation function for all classifiers. Although titles and descriptions are textual data, the authors chosen to treat them as different modalities, which allowed different pre-processing steps. The authors identified that each product belongs, on average, to 3 classes.

For description pre-processing, authors removed stopwords, whitespace, numeric digits, punctuation, and words longer than 30 characters, and truncated descriptions to be less than 300 words long. In titles, sentences were limited to 57 words and they didn't remove stopwords. Then, they used Glove to get the word embeddings that were later used for classification. In the classification, the authors used a modification of the CNN (KIM, 2014), both for the title and for the description model. For the visual model, they used ResNet-50 pre-trained in ImageNet (DENG et al., 2009) and available in the Keras API, without the classification layer and with a fully connected layer with 122 parameters

[9] The code for this work is not available.

added to correspond to the number of labels. Table 8 shows the values obtained for the evaluation of unimodal models.

Table 8 – F1 values for each unimodal classifier.

Modal	F1(%)
Image	61.0
Title	**82.7**
Description	77.0

Source: (WIROJWATANAKUL; WANGPERAWONG, 2019).

For the fusion of these unimodal models, the authors tested Simple Linear Regression, Bi-Modal Fusion and the proposed model, the Tri-modal Fusion. For Linear Regression, the authors trained a model for simply merging the individual classifiers into a single one. In Bi-Modal fusion, combinations of two neural networks were tested – one for title and one for image – for purposes of comparison with the late fusion approach, where they merged the output of the two classifiers. The authors also merged title and description with two networks; and also tested another fusion involving a network for description joined with another one for image. Finally, the proposed model, which merges the three modalities at the decision level, was evaluated. Table 9 shows the results.

Table 9 – F1 values for each multimodal combination.

Model	F1(%)
Max	78.8
Mean	81.7
Linear Regression	83.0
Image-Description Fused	82.0
Image-Title Fused	85.0
Title-Description Fused	87.0
Image-Description-Title Fused	**88.2**

Source: (WIROJWATANAKUL; WANGPERAWONG, 2019).

Although they obtained good F1 values, the authors say that some categories had a significant number of errors (Table 10) and that they would like to better explore with a greater amount of data in future work.

This work also guided our option to use late fusion, and in its multimodal (tri-modal) approach, the authors use title and description separately in CNN , whereas in our work, title and description go together (early fusion) as just one textual modality. Finally, this work also used WE produced by Glove (122-dim) while we used FastText (64-dim) trained on specific-domain corpus.

Table 10 – The top-15 most misclassified categories using the tri-modal fusion approach.

Fused Image-Title-Description	
1 Chew Toys	(64/132)
2 Accessories	(417/924)
3 Women	(58/134)
4 Novelty, Costumes & More	(127/303)
5 Snacks	(143/363)
6 Hunting Knives	(59/152)
7 Men	(112/291)
8 Clothing, Shoes & Jewelry	(256/686)
9 Hunting & Tactical Knives	(65/175)
10 Shampoos	(58/158)
11 Balls	(67/185)
12 Squeak Toys	(79/223)
13 Horses	(92/265)
14 Boating	(113/327)
15 Airsoft	(68/200)

Source: (WIROJWATANAKUL; WANGPERAWONG, 2019).

3.4 Is a Picture Worth a Thousand Words? A Deep Multi-Modal Architecture for Product Classification in E-Commerce (ZAHAVY et al., 2018)

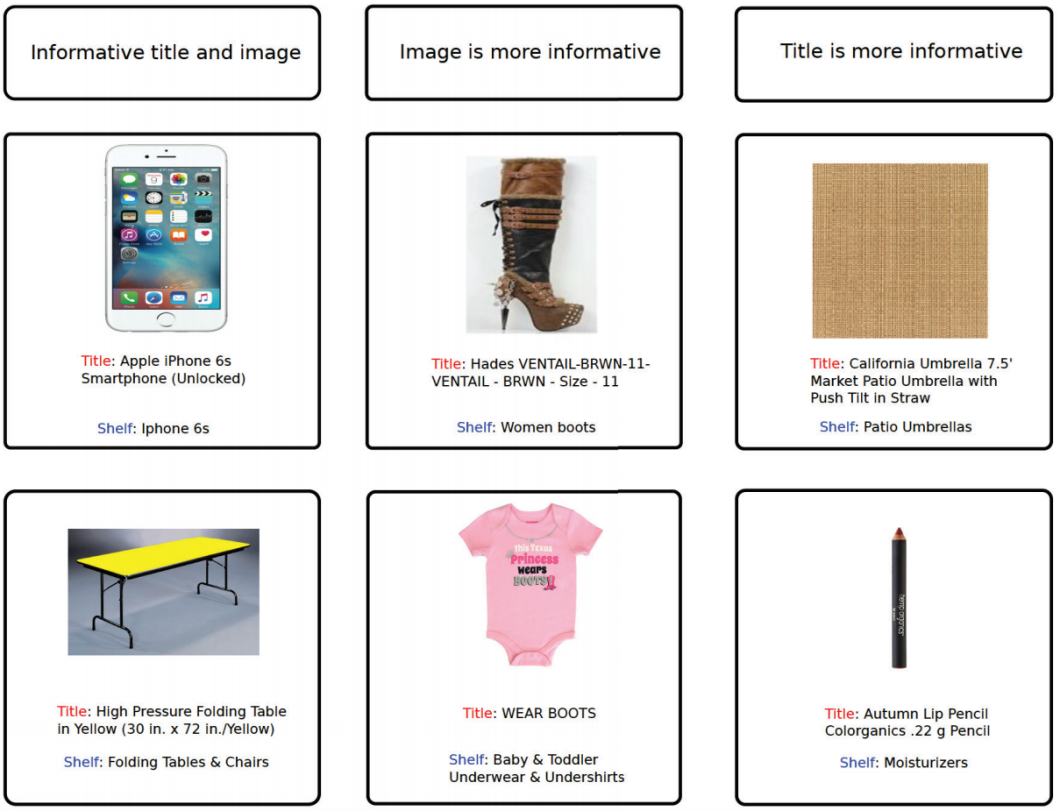
Zahavy et al. (2018) propose a deep neural network for multimodal classification of products where the final architecture has three main components: (1) a CNN of (KIM, 2014) for the text, (2) a CNN of (SIMONYAN; ZISSERMAN, 2015) for image, and (3) a decision neural network that learns to choose the classification between these two modalities. For this work, they collected a large-scale dataset of 1.2 million products on the website Walmart.com[10]. Each product has a title, an image and needs to be classified on a label (shelf) from 2,890 possible labels. Examples of this dataset can be seen in Figure 14 and are also available online at Walmart.com.[11]

For most products, the image and title contain information relevant to customers. However, according to Zahavy et al. (2018), for some of the products, both types of input may not be informative for the shelf forecast (see Figure 14 for examples). For the authors, this observation motivated their work and raised questions such as: which type of input is most useful for classifying products? Is it possible to merge the inputs to reach a better architecture?

In an attempt to answer these questions, the authors proposed the method shown in Figure 15. In text mode, they use the CNN of (KIM, 2014) where the first layer represents random words (different from the original paper). The next layer performs convolutions

[10] <Walmart.com>
[11] The code for this work is not available.

Figure 14 – Walmart data in which the leftmost column shows products in which the text and image are equally informative. In the middle column, there are products in which the image is more informative and finally, the column on the right has texts which are more informative than the images.



Source: (ZAHAVY et al., 2018).

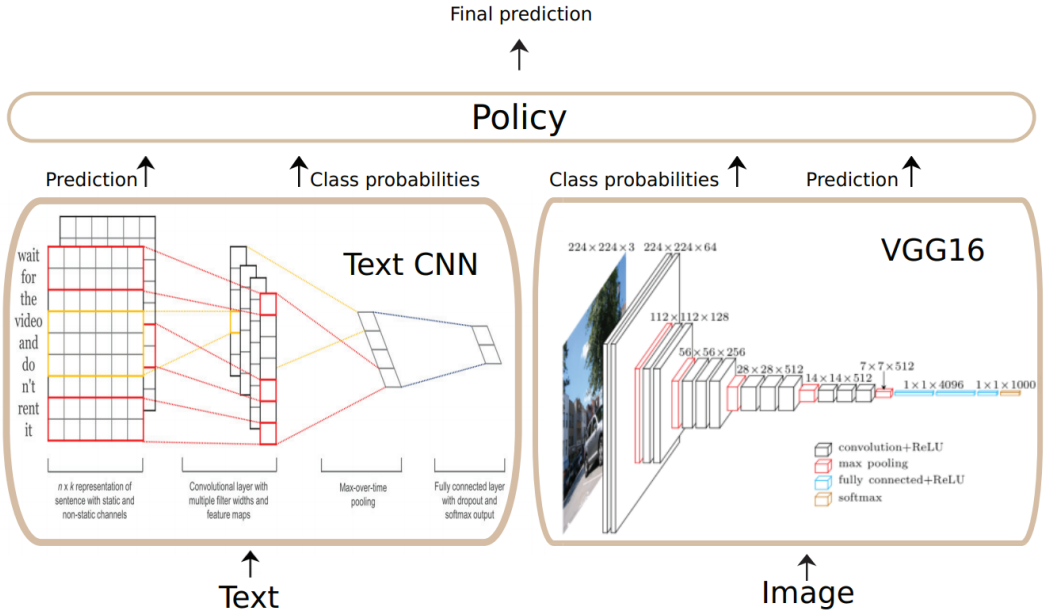
over time on the word embeddings using various filter sizes (3, 4 and 5). Then the max-pool-overtime step concatenates each convolution filter together with the results. In this module, the authors added a regularization layer of *dropout* (drop rate of 0.5), followed by a fully connected layer, which will be classified by a *softmax* activation layer.

For image classification, they used the VGG network (SIMONYAN; ZISSERMAN, 2015). The input to this network is a fixed-size RGB image of 224 x 224 pixels, which passes through a stack of convolutional layers that is followed by three dense or highly connected layers. The first two fully connected layers have 4,096 parameters each, the third performs product classification of 2,890 classes and therefore contains 2,890 parameters (one for each class).

The authors concluded that, for classification, the experiments suggest that text information is more informative than images and they also demonstrated that when learning a decision rule with a deep neural network it is possible to obtain a better performance than with unimodal architectures.

As in our research, this work used CNN for the visual modality, but in that case it was VGG16 while we chose Inception Resnet V2. They used a CNN adapted for text and we use CNN above our domain-specific FastText WE.

Figure 15 – The decision-level (late) multimodal fusion architecture proposed by Zahavy et al. (2018)



Source: (ZAHAVY et al., 2018).

3.5 Final considerations

In Table 11 the main characteristics of the works cited in this chapter are presented regarding: (i) the dataset used, (ii) the task for which the method was proposed, and (iii) the approach and level of fusion that the method used.

Table 11 – Works cited in this Chapter 3 and their main characteristics.

Work (section)	**Dataset**	**Approach used**	**Fusion level**
3.1	Rakuten France: Multimodal Product Dataset (99 thousand products)	CamemBert, ResNet152, DCL Method	Early and late-fusion
3.2	Rakuten France: Multimodal Product Dataset (99 thousand products)	CamemBert, ResNet152, Co-attention; Bi-Directional Long Short Term Memory(BILSTM)	Late-fusion
3.3	9,4 thousand products from Amazon	two CNNs, a ResNet50 and a Feed-forward Neural Network	Late-fusion
3.4	1,2 million products from Walmart.com	VGG16, CNN for textual modality and another to final prediction	Early and late-fusion

Source: The author.

Our experiments, unlike the works above, used a dataset (`compliance-small`) that has only 3 classes and few products (4,334 for each class), whereas the other works experimented with datasets with many more classes and products. Our approach used WE trained using FastText in the e-commerce domain (ROMUALDO; REAL; CASELI, 2021), while some of the related works used contextualized language models such as general domain or multilingual ones. We used Inception ResNet V2, while the other works used ResNet 152 or VGG16. We do early-fusion in our TD (Title and Description) model for WE extracted by title and description separately, while other authors did something similar either using embeddings for all text or late-fusion for their textual models. For predictions we used CNN.

Chapter 4

Multimodal classification of non-compliant products

This chapter discusses the problem addressed in this work, as well as the materials and methodology employed.

4.1 Problem description

In a marketplace, a product is typically represented by metadata, such as title, descriptions and images, with the majority of this information being manually assigned by the seller. In general, a marketplace allows sellers to automatically enter any product into the sales platform. This flexibility is required to cover the wide range and quantity of products that form the assortment of a marketplace. However, it can lead to legal issues if a seller places a product on the sales platform that cannot be sold there. Some products may be sold in certain countries but not in others; some sellers may be authorized to sell or resell specific branded products while others are not; and each marketplace has its own policies that may prohibit the sale of certain items. As a result, each marketplace has distinct compliance requirements to ensure a fair assortment.

Companies use a variety of methods to ensure that their assortment is as accurate as possible, including implementing procedural business rules, manually checking, and automatically detecting non-compliant products. In this work, we look into the case of Americanas S.A., Latin America's largest marketplace with over 200 million active offers. Given the complexities and scope of our scenario, we looked into the efficacy of machine learning approaches for detecting products classified as Adult or Illegal Devices. The classification problem of these products was treated as a binary classification problem at

first and also as a multiclass one. The classes in these cases were, respectively: Adult (Positive or 1), Illegal Devices (Positive or 2) or not (Negative or 0).

Unimodal classification can be done using just one modality of each instance, such as textual information from title and description of a product (textual modality) or their images (visual modality). On the other hand, multimodal classification is performed based on both information: textual and visual. Thus, the research question risen in this work is: is there an inherent feature of textual or visual modality that allows the classification of a product as Adult or as Illegal Device **or** the accurate classification of it is only possible with the combination of features from different modalities?

In this chapter we present the materials and methods applied to answer this research question. The results of the classification of products as Adult or Illegal Devices are presented in Chapter 6.

4.2 Materials

The data provided by Americanas S.A. can be divided into three sets: (1) e-commerce-large, (2) e-commerce-small, (3) compliance-large and (4) compliance-small. The large e-commerce data contains approximately 7.5 million of textual information about products which, combined with titles and descriptions, total approximately 8 billion words. This set was used to generate textual representations (word embeddings) using several different techniques such as Word2Vec, FastText and Glove as detailed in (ROMUALDO; REAL; CASELI, 2021). The second dataset, e-commerce-smaller, contains title, description, and image information for nearly $3,000$ products. It has also a matching field which indicates the correspondence between similar products. The first two datasets – e-commerce-large and e-commerce-smaller – were used in multimodal similarity scoring experiments from (ROMUALDO; REAL; CASELI, 2021). These experiments will be not detailed in this dissertation but they were useful to chose the word embeddings used in the multimodal classification problem addressed this work.

There are various classes of products in the set compliance-large, ranging from technological appliances to kitchen utensils, furniture, or even non-compliant products, which are products that cannot be sold on the marketplace due to the partner company's sales policies. Each of these items has a title, a textual description and an image associated to it. The number of images associated to each item and the accuracy of the information are determined by the seller who published them. Among all the product classes in this set, 19 of them encompass items that are prohibited from being sold on the marketplace, such as: adult content, e-cigs (as electronic cigarettes), firearms, medical devices, illegal substances such as drugs, pharmaceuticals, among others. The compliance-large set has some uncertainties regarding the label (class) assigned, as some class are duplicated. The duplicated class include: pharmaceuticals and pharmaceuticals, or electronic cigarettes

and electronic cigarettes. The approach carried out to solve this problem is explained later in this document.

For this work, a compliance-small set was built to carry out the experiments presented in Chapter 6. The data used in the classification experiments presented in this work refer to 15,565 products (compliance-small), 6,000 of them from **Negative** class (products allowed to be sold) and the remainder composed of non-compliant products, with 4,334 of them belonging to the **Adult** class and another 5,231 belonging to the **Illegal Devices** class. The 6,000 Negative class products were chosen from a larger set of 213,454 products available on the Americanas S.A.'s marketplace.

The following steps were taken to select the items for the Negative class: (i) removing duplicated products (which reduced the initial set of 213,454 products to 63,780 products); (ii) clustering of the remainig products using the K-means algorithm to generate 30 clusters; and (iii) selecting 200 products from each cluster, with half (100 products) randomly selected as those within the mean distance from the centroid and the other half (plus 100 products) randomly selected as those above the mean distance from the centroid. At the end of this process, 6,000 products were chosen for the Negative class. To keep the classes balanced, the number of products was limited to the number of data in the smallest class, which 4,334 in each. For the K-means algorithm, K was empirically defined as 30 by the Elbow method with mini-batch k-means, which tests which is the best value of K. For this grouping we used the title obtained after pre-processing with the removal of special characters, lowercase and stopwords.

Thus, the compliance-small dataset contains the following data in addition to the images directory: product id, product name, product description, and label (Negative, Adult or Illegal Device). The image was retrieved based on the id column's content, which corresponds to the name of the image file in that local directory.

4.3 Methods

The general flow for all models tested in this work is represented in Figure 16. We experimented with 15 models: 5 generated for Adult × Negative classes, 5 generated for Illegal Devices × Negative classes, and 5 generated for multiclass classification with three classes – Negative, Adult, and Illegal Devices. As shown in the diagram, classification entails several modules that are important for the classification task in this work, where several strategies were investigated:

WE Model – *Word embeddings* are one of the many possibilities investigated in the literature. In this case, models were created taking into account only information from titles (title models), only information from descriptions (description models), or both (title+description models). To chose the best WE model for our experiments,

we trained several domain-specific e-commerce language models (using e-commerce-large set) and compared them with domain-general models in a similarity calculation task (based on e-commerce-small set), as detailed in (ROMUALDO; REAL; CASELI, 2021). Based on the results of these experiments, we chose the 64-dim FastText (BOJANOWSKI et al., 2017) model generated following the SkipGram approach.

CNN model – The reuse of previously models such as ResNet, VGG, and others is one of the most commonly used approaches in literature. Unlike texts, the reuse of domain-general models in a specific domain appears to have less impact, but this assumption has yet to be tested. Based on this assumption, in the experiments presented in this work we chose to optimize feature extraction and learning from pre-trained Inception ResNet V2 model (SZEGEDY et al., 2017). We use transfer learning from ImageNet and fine-tuning to our models.

Dense Classifiers – In this step, classification variation occurs due to activation settings, such as softmax for multiclass and sigmoid for binary classification, as well as training with the appropriate product classes. For the binary classification, fifteen models were created: five considering only the Adult class, five considering only Illegal Devices, and five that combined the two classes in a multiclass configuration. Among these 15 models are ensemble models, which combine the individual pre-trained models to improve the detection of the non-compliant products.

The models generated following the flow described in Figure 16 are referenced by these letters: Only title (`T`), Title and Description (`TD`), Visual (`V`), Multimodal (`M`) and Ensemble (`E`). All of them are described in details in the next subsections.

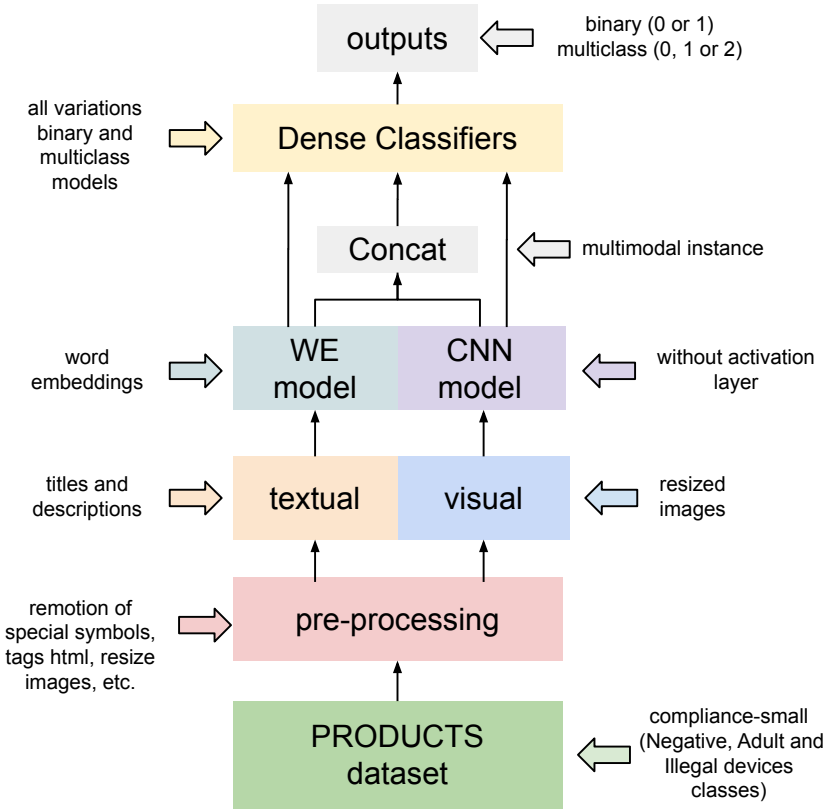
Regarding the Products dataset – the input for the model generation flow of Figure 16 – there are different dataset splits, according to each model, as shown in Table 12. Please note that for testing purposes, all models have the same testing percentage representing the same data.

Table 12 – Splits of the dataset in train, validation and test sets, for each model.

Model	**Train**	**Validation**	**Test**
Only title (`T`)	50%	20%	30%
Title and description (`TD`)	50%	20%	30%
Visual (`V`)	50%	20%	30%
Multimodal (`M`)	40%	30%	30%
Ensemble (`E`)	50%	20%	30%

We use 40% for Multimodal training (`M`) because this model uses transfer learning from other models, which accelerates the model learning. Thus, this model was used for training with some layers trainable (fine-tuning). For example, the layers closest to the input layer are frozen and the last layers are trainable (adjusted in fine-tuning).

Figure 16 – General unimodal and multimodal processing flow followed in this research for the binary and multiclass classification of non-compliant products.



Source: The author.

4.3.1 Textual models

For the inputs of the textual models, *word embeddings* generated by FastText (BOJANOWSKI et al., 2017) using the SkipGram approach and with 64 dimensions were used. This textual representation was chosen based on the experiments detailed in (ROMUALDO; REAL; CASELI, 2021).

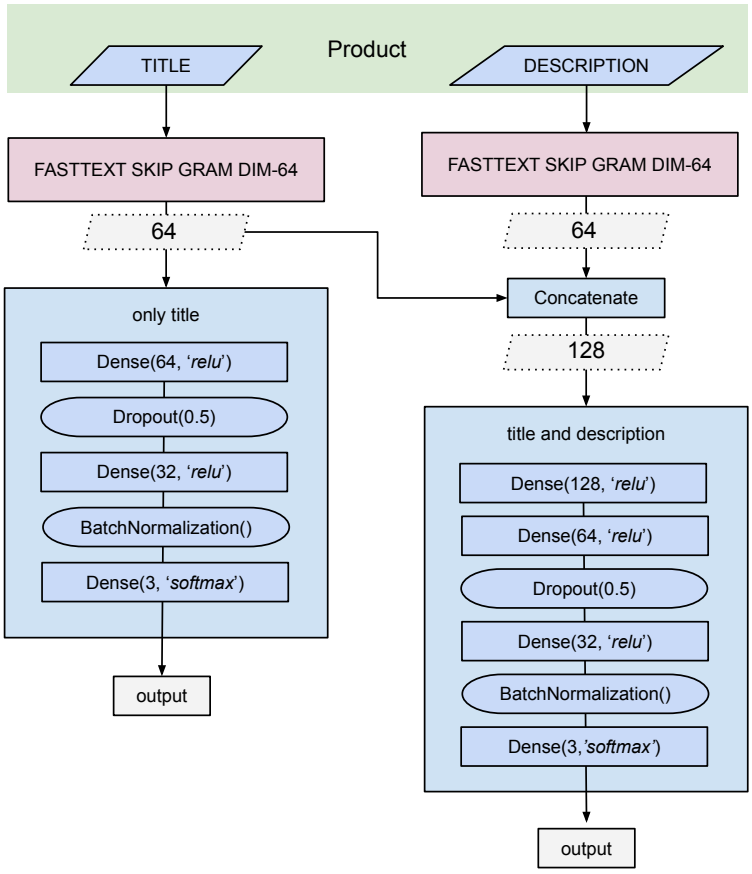
Before getting *word embeddings* by FastText {3}, all texts {1} are preprocessed in order to remove *stopwords*, special characters, numeric sequences, links and perform the conversion to lowercase {2}, as illustrated in the example:

```
1. "Faca Esportiva Xingu XV2562 Outdoor com Bainha eBússola Camuflada"
2. "faca esportiva xingu xv outdoor bainha bussola camuflada"
3. "[ 0.01715468, -0.01149213, 0.03243327, ... ,-0.06800657, -0.03518752]"
```

Two textual models were generated in the experiments to classify items in the Adult and Illegal Devices categories: one for titles only (`T`) and another for titles and descriptions (`TD`). The input to the `T` model is a 64-dimensional vector of the title's *embeddings*. The input to the `TD` model is two concatenated vectors: the first one for the title and the second one for the description, totaling 128 as shown in Figure 17. Keras (CHOLLET et

al., 2015) was used to implement all models. The Binary Cross Entropy loss was used for binary models, while the Categorical Cross Entropy loss was used for multiclass models. The binary models' activation layers were sigmoid, and the multiclass softmax model's activation layers were softmax.

Figure 17 – Description of textual unimodal models. Note that all templates use the 64-dimensional *words embeddings* FastText SkipGram. In binary models, the value of the last Dense layer (activity layer) is reduced from 3 to 1.



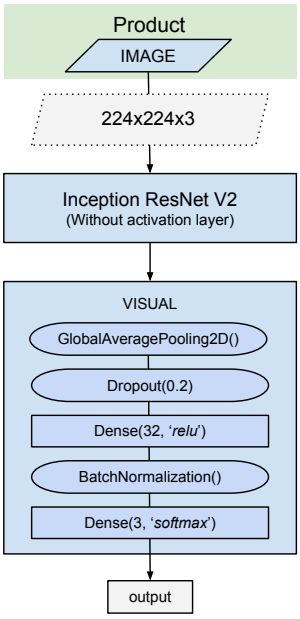
Source: The author.

The training was performed with 100 epochs, *batch* size equal to 32 and an Adam learning rate of 1×10^{-5}, with a total of 6,849 and 27,009 parameters for the models of binary T model and TD model and for multiclassing are 6,883 parameters for T model and 27,075 for TD model. The machine used for training the models has Windows 10 build 20H2, AMD Ryzen 5 3600 6-Core Processor with 12 cores up to 3.6GHz, 32GB of RAM and an NVIDIA GeForce RTX2060 video card with 6GB of dedicated VRAM. Each model was trained for around 2 minutes.

4.3.2 Visual models

In the visual model (`V`), as showed in Figure 18, all images were resized to 224×224 pixels, keeping the dimensions that represent the RGB colors.

Figure 18 – Illustration of the combination of the pre-trained Inception ResNet V2 model the value of the last Dense layer (activity layer) is reduced from 3 to 1.



Source: The author.

For this visual model, we also used Keras (CHOLLET et al., 2015), which has several models of convolutional neural networks (CNN) pre-trained in datasets such as ImageNet (DENG et al., 2009), MNIST (LECUN; CORTES, 2010), CIFAR (KRIZHEVSKY, 2012) and others. To build this model, techniques such as *transfer learning* and *fine-tuning* were used to optimize feature extraction and learning.

For *transfer learning*, we used the Inception ResNet V2 (SZEGEDY et al., 2017) network pre-trained for ImageNet with 1,000 classes. Its architecture has 780 layers and the classification layer was removed and concatenated to the visual model of this experiment. For *fine-tuning*, a freeze is performed from the initial layer to layer 750, in order to ensure that the model performs the same feature extraction as in its pre-training. The visual model (Figure 18) generated in this way makes the classification based on the *features* of an image extracted by the Inception ResNet V2 network. This was an attempt to preserve the unimodal features already learned by the original model.

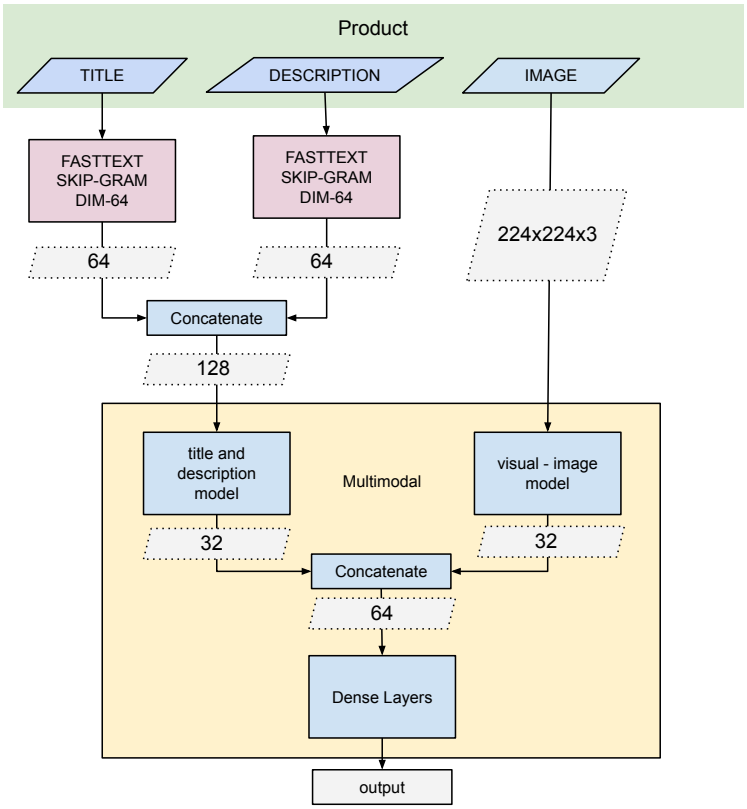
The training was performed with 30 epochs, a *batch* size of 8 and an Adam learning rate of 1×10^{-6}, with a total of $54,386,147$ parameters for the multiclass and $54,386,081$ for those with binary classification. The machine used to train this visual model was the same as the one used for the textual model and the training takes about 30 minutes.

4.3.3 Multimodal models

The multimodal model, illustrated in Figure 19, uses pre-trained unimodal models without their activation layers. Models are concatenated and then transported across dense layers. The activation function is also *sigmoid* when the classification is binary and

softmax for multiclass, also the value of the last Dense layer (activity layer) is reduced from 3 to 1, the output represents the final classification of the model. To use the model for the binary classification, it can be considered, for example, that a value of y greater than 0.5 indicates a product that should be considered "adult" or "illegal device" (output/class equals to 1) while an y value lower or equal than 0.5 indicates a product that can be sold (output/class equals 0). In a multiclass scenario, the output is a 3 position vector where the highest value position represents the detected class, Negative (0), Adult (1) or Illegal device (2).

Figure 19 – Description of the textual-visual multimodal model produced in this work.



Source: The author.

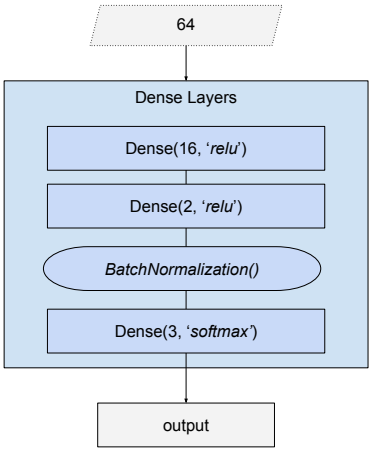
The figure 20 details the dense layers of the multimodal model.

For the training 30 epochs, a batch size equal to 8 and an Adam learning rate of 1×10^{-4} were set. In this multimodal model, The visual module is set to be untrainable until layer 650, and the textual module of description and title is set to be untrainable until the *layer* "Dense(32, 'relu')" (see Figure 17). Again, the same machine was used for training the multimodal model for take about around 35 minutes.

4.3.4 Ensemble models

The ensemble (`E`) models use all unimodal and multimodal pre-trained models, with their activation layers, getting their output values as illustrated in Figure 21. In total

Figure 20 – The dense layers of the multimodal model are specified in Figure. 19. In the binary model the *softmax* activation function is changed to *sigmoid* and the value of the last Dense layer (activity layer) is reduced from 3 to 1, where the output represents the final classification of the model.

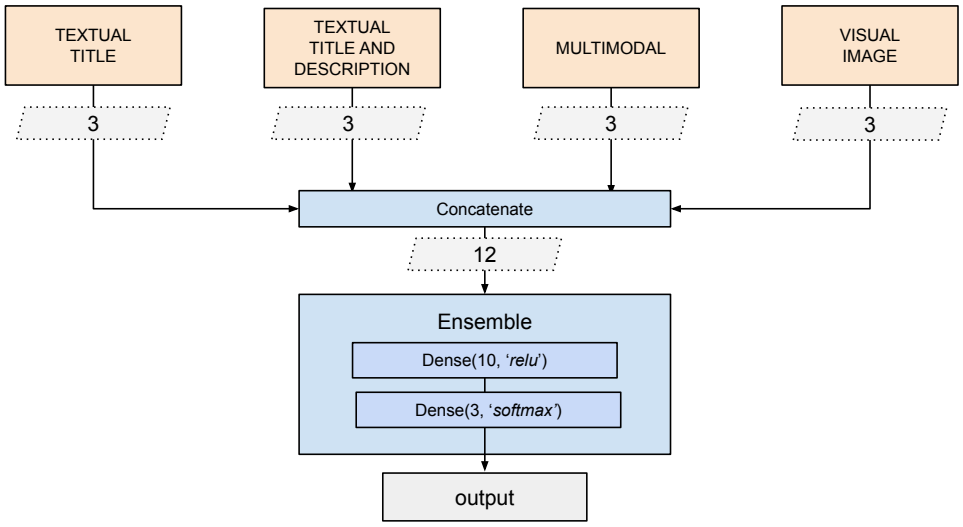


Source: The author.

of 3 ensemble models were generated. The models are concatenated after the activation layer and transported by a dense layer classifier of the ensemble where they are the only ones to be trained. The activation function for multiclass models is the *softmax* and for binary models is the *sigmoid* and we set 30 epochs, *batch* size equal to 16, *loss categorical cross-entropy* and an Adam learning rate of 1×10^{-4}. With the 108,836,495 parameters and $108,836,221$ for the binary classification models, the train of the multimodal model took about 1 hour in the same machine used to train the unimodal models.

The number of layers and structure of this model was defined based on empirical test of the joining of the binary models that occurred before the multiclass models. The outputs of the models were joined and trained with the classification between them.

Figure 21 – Description of the ensemble where all 4 models are concatenated.This image represents multiclass model and the activation function is *softmax*, and *sigmoid* for binary classification. Also the value of the last Dense layer (activation layer) is reduced from 3 to 1, and the output represents the final classification of the model.



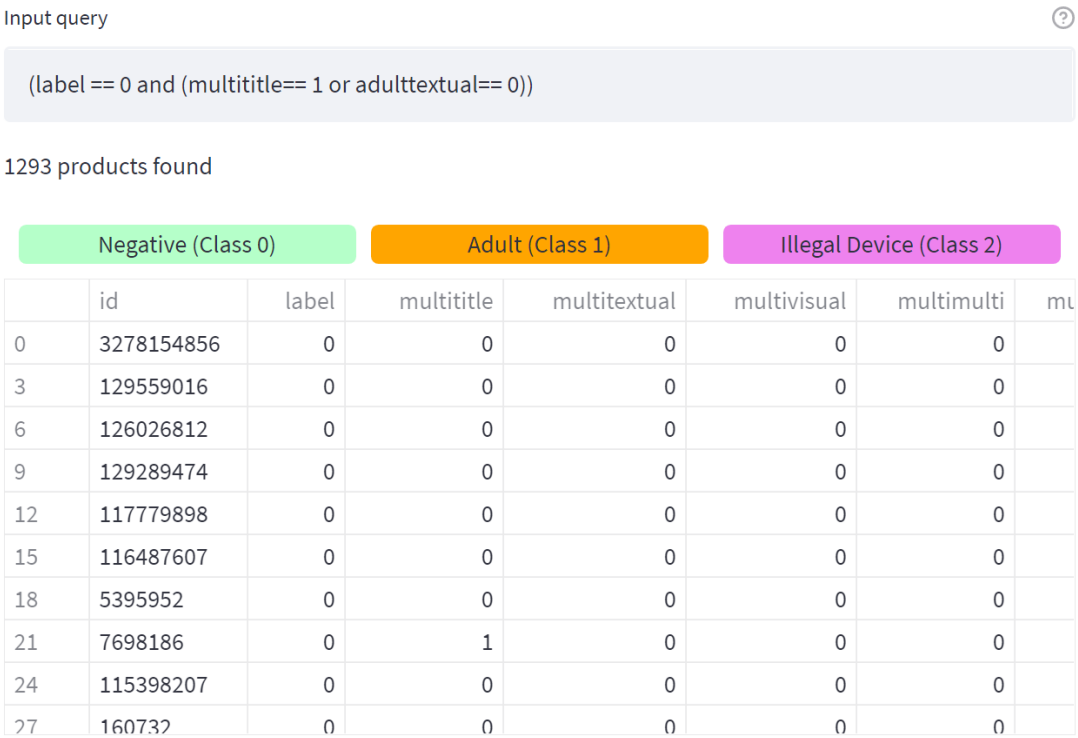
Source: The author.

Chapter 5

Error analysis tool

To facilitate the qualitative analysis of the results of the experiments described in Chapter 6, an error analysis tool, illustrated in Figure 22, was implemented to analyze the classification errors in detail.

Figure 22 – Main screen of the error analysis tool.



	id	label	multititle	multitextual	multivisual	multimulti	mu
0	3278154856	0	0	0	0	0	
3	129559016	0	0	0	0	0	
6	126026812	0	0	0	0	0	
9	129289474	0	0	0	0	0	
12	117779898	0	0	0	0	0	
15	116487607	0	0	0	0	0	
18	5395952	0	0	0	0	0	
21	7698186	0	1	0	0	0	
24	115398207	0	0	0	0	0	
27	160732	0	0	0	0	0	

Source: The author.

By means of an input query, the tool is able to search for instances considering different classification configurations. As shown in Figure 22, there is a space for specifying the

query indicating the desired value for the correct `label`, as well as for the outputs returned by the models being evaluated (in this case, the `multititle` and `adulttextual`). In the example in this figure, *query* queries for products that can be sold (Negative class, `label == 0`) but have been misclassified as Adult by the title-based multiclass model (`multititle == 1`) and correctly by the adult binary model based on title and description (`adulttextual == 0`). In this image it is possible to observe the *ID*s of products with all the classifications output by the models, and these IDs can also be used as a search in the *query*.

If one wants to recover the Negative products (`label == 0`) that were correctly classified as Negative by any binary model trained considering only the products from Adult class (`(adulttitle==0 or adulttextual==0 or adultvisual ==0 or adultmulti == 0 or adult ensemble==0)`) but were classifiied as Illegal Devices by the multiclass multimodal model (`multimulti==2`), the *query* should be:

```
label==0 and (adulttitle==0 or adulttextual==0 or adultvisual==0 or
adultmulti==0 or adultensemble==0) and multimulti==2)
```

Another possible *query* is: which Illegal Devices (`label==2`) does the title model trained using only the illegal devices' titles classified as Negative (`disptitle==0`) while the multiclass model trained based on only the titles classified as Adult (`multititle==1`)? In this case, such a query is possible through the query:

```
label==2 and disptitle==0 and multititle==1
```

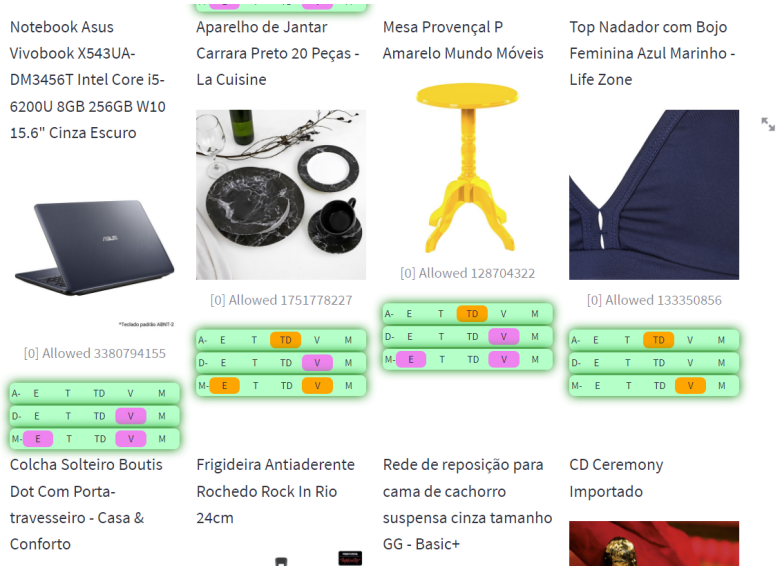
It is also possible to check the n-grams present in the textual values associated with the products (as illustrated in Figure 23), by selecting a value N that varies from 1 (unigram) to 8 (8-gram). In this *query*, the top-k n-grams are returned where k ranges from 10 to 50 and the top-k is defined considering a descending order of frequency ranking. This n-grams *query* returns 3 graphs representing the n-grams present in: (1) title+description, (2) title only and (3) description only, as shown in Figure 23.

Finally, in Figure 24 we can see some products resulting from a *query*. Each product is accompanied by its title, its image (photo) and its correct class (in this case, all the products are from Negative class). Below this product information, there are three bars representing the classifications output by each of the 15 trained models. These bars indicate, from top to bottom: the binary classification models trained from Adult class data (`A`) or Illegal Devices data (`D`), and the multiclass classification models (`M`). Each model configuration is also indicated as: `E` (ensemble), `T` (textual model generated only from product titles), `TD` (textual model generated from product titles and descriptions), `V` (visual model) and `M` (mutimodal model). The colors indicate the correct classes, being:

Figure 23 – n-grams investigation functionality of the error analysis tool

Source: The author.

Figure 24 – Product classification in the error analysis tool.



Source: The author.

green for Negative (a compliant product), orange for Adult and pink for Illegal Devices.

By means of this interface, it is possible to check which model wrongly classified each instance. Furthermore, it is also possible to verify which product were classified in the same way by different models, as well as for which items the classification was different or correct (an adaptation of the confusion matrix to visualize the products).

For example, let's consider the notebook example on the left of the image. In this case, 3 models incorrectly classified the product as Illegal Device: (1) the binary visual model

(`V`) trained considering only Illegal Devices (`D`), (2) the ensemble model (`E`) of multiclass classification (`M`) trained considering all classes and (3) the visual model (`V`) of multiclass classification (`M`) trained considering all classes.

The yellow (*amarelo*) table (*mesa*), in turn, was misclassified as an Adult product by the binary title-description textual model (`TD`) trained considering only Adult products (`A`), and as Illegal Devices by other three models (the same that misclassified the notebook).

Chapter 6

Experiments and Results

In this chapter, we describe the results of the experiments carried out with the classification models trained for binary and multiclass classification tasks as detailed in Chapter 4.

As detailed in Section 4.3, 15 models were generated in this work: 5 generated for Adult × Negative classes, 5 generated for Illegal Devices × Negative classes, and 5 generated for multiclass classification with three classes – Negative, Adult, and Illegal Devices.

Quantitative and qualitative results are presented separately in the following sections.

6.1 Quantitative results

The values of precision, recall and $F1$ (see Section 2.3) for the unimodal (`T`, `TD`, `V`), multimodal (`M`) and ensemble (`E`) models, for binary and multiclass classification are presented in Tables 13 and 14, respectively. We also compared our results of multiclass classification (Table 14) with the ones obtained in a binary scenario where we considered the products that can be sold as Negative and the Adult plus Illegal Devices as Positive.

It is important to say that in our task of classifying non-compliant products, it is preferable to have a high recall for positive classes (Adult or Illegal Devices), as not detecting them is worse than classifying a Negative product as Adult or Illegal. Thus, all the conclusions presented in this chapter were taken on the basis of higher recall values.

In Table 13 we can see the summary of the results for the binary models trained for detecting Adult products and Illegal Devices. The multimodal (`M`), title + description (`TD`) and ensemble (`E`) models shown the best recall values (in bold), for the Adult products, with just one percentage point differing the ensemble (the best) model from the other two. For the Illegal Devices, in turn, the best models were the multimodal ones, `M` and `E`

with 96% of recall for the Positive class but again with only one percentage point above the `TD` model.

Table 13 – Binary classification of Adult or Illegal Device (class 1) products against the Negative (0) class.

		Adult (`bin-adult`)			**Illegal Device (`bin-illegal`)**		
model	**class**	**precision**	**recall**	$F1$	**precision**	**recall**	$F1$
`T`	0	97%	98%	97%	95%	99%	97%
	1	98%	97%	97%	99%	94%	97%
`TD`	0	98%	100%	99%	96%	100%	98%
	1	100%	**98%**	99%	100%	**95%**	97%
`V`	0	87%	84%	96%	86%	89%	88%
	1	85%	87%	86%	89%	86%	87%
`M`	0	98%	99%	99%	96%	99%	98%
	1	99%	**98%**	99%	99%	**96%**	98%
`E`	0	99%	99%	99%	96%	100%	98%
	1	99%	**99%**	99%	100%	**96%**	98%

For the multiclass classification, the best model regarding the highest overall recall values for the Positive classes was the multimodal (`M`). However, again, the `TD` and `E` models are very close. In general, the multiclass approach did not outperform the binary one.

Table 14 – Multiclass classification of Negative (0), Adult (1) and Illegal Device (2) products.

model	**class**	**precision**	**recall**	$F1$
`T`	0	97%	97%	97%
	1	94%	96%	95%
	2	99%	**96%**	97%
`TD`	0	98%	**99%**	99%
	1	94%	**98%**	96%
	2	100%	94%	97%
`V`	0	73%	83%	78%
	1	83%	78%	80%
	2	82%	75%	78%
`M`	0	98%	**99%**	99%
	1	95%	**98%**	96%
	2	99%	**95%**	97%
`E`	0	96%	86%	91%
	1	89%	**99%**	94%
	2	92%	92%	92%

The worst performance of visual models in both tasks (binary and multiclass classification) indicates, at very least, that better models seem to be needed to give them a chance against the textual ones. However, it is worth mentioning that the unimodal models trained for binary classification, when merged as multimodal or ensemble models,

led to a little increase in some measure values, indicating that visual information can, indeed, complement the textual one which provides clues for the hypothesis of this work.

Furthermore, it is important to point out that in a production scenario, the `TD` models have a greater advantage since they were more agile in training and their time to predict the class of a given product was approximately 0.5 seconds for almost 4,000 products. Ensemble models took about 3.5 minutes to predict the same set of products.

Thus, the best configuration for classifying non-compliant products according to the quantitative results of the experiments carried out in this work is the combination of two binary models: the **`TD bin-adult`** (with a recall of 98%, see Table 13) and **`TD bin-illegal`** (with a recall of 95%, see Table 13). The `E bin-adult` model, with a recall of 99% and the `E bin-illegal` with a recall of 95% (see Table 13) have a bigger prediction cost, so they were passed over by the `TD` versions. The lower prediction cost also guided our choice for the `TD` models in spite of the `M` ones.

6.1.1 Discussion

In binary classification models, it is believed that the textual models may have focused on specific words and, thus, have caused the correct classification of products. In terms of visual models, the variation of images of different types of products in the category may have brought noise to the learning process.

The Adult class also has several products, such as clothes and other items, which even for humans can be confusing to classify. Depending on the way the product is presented, it may or may not be considered an Adult product by the viewer. Also, it is possible that all the images of the classes get into a kind of spatial conflict, in which the model is not able to correctly detect the class among the images, due to the pattern of this category be more conceptual than some visually detectable pattern in the images. Although Illegal Devices are possibly closer to other electronic devices, such as cameras, TV controls etc., the model can confuse and lead to a misclassification by some format in the image, placing it in the Adult class, for example, due to the learning of the patterns (perhaps unknown) of the classes involved.

However, visual models are asked to deal with a very difficult task. Maybe, if we add a significant number of class-negative products, the models could be able to capture the characteristics of what our illegal or adult products really are. That's because the positive set (adult products or illegal devices) is just one set among the whole universe of product categories that the model has to deal with.

For the specific task of classifying Adult products, the visual information seems not to have a big impact in comparison to the textual one. Maybe, for this kind of task and product, the textual information give more clues about the non-compliance feature of an Adult product than the visual one. On the other hand, the combination of visual and textual information seems to benefit the identification of Illegal Devices, since the two best

models were the ones trained from visual and textual information: the multimodal and the ensemble. However, the little difference (just 1 percentage point) between `E` and `TD` model, and the little prediction time of `TD` model made us choose the TD binary models as the best ones.

Thus, unfortunately, the answer for our research question – "is there an inherent feature of textual or visual modality that allows the classification of a product as Adult or as Illegal Device **or** the accurate classification of it is only possible with the combination of features from different modalities?" – is not straight. There are lots of factors to be considered to answer this question. However, from our experiments, we can conclude that the textual information has a big positive impact in the performance of the classification models trained in this work.

6.2 Qualitative results

To understand the behavior of the models, we analyzed some of the test instances from Table 15, with special attention to those that were classified differently by the trained models. In this table, the columns indicate the trained model – title (T), title and description (TD), visual (V), multimodal (M) and ensemble (E) – subdivided for multiclass classification (M), Adult class (A) or Illegal Devices (I) binary classification. The highlighted lines represent the product classes that were classified correctly, followed by the other classes without highlighting. Thus, from this table, it is possible to analyze the distribution of products classified by the models.

The binary classification models trained from Illegal Devices data, except the visual (`V`), were the ones with the most correct classifications for the Negative class. The multiclass ensemble (`E`) correctly classified 5 more products from the Adult class than the binary classification title and description (`TD`) model did. For the class of Illegal Devices, the best model was the multimodal binary classification.

Table 15 – Confusion matrices for the trained models. The highlighted lines represent the product classes that were classified correctly, followed by the other classes without highlighting.

	T			**TD**			**V**			**M**			**E**		
	M	A	I	M	A	I	M	A	I	M	A	I	M	A	I
Negative	1259	1264	**1293**	1280	1284	**1295**	1083	1095	1115	1287	1289	**1290**	1120	1287	**1295**
Adult	33	35	-	15	15	-	108	204	-	6	10	-	83	12	-
Illegal Device	7	-	6	4	-	4	108	-	144	6	-	9	96	-	4
Negative	41	31	-	17	19	-	177	165	-	24	20	-	12	20	-
Adult	1251	1268	-	**1280**	**1280**	-	1011	1134	-	1273	1279	-	**1285**	1279	-
Illegal Device	7	-	-	2	-	-	111	-	-	2	-	-	2	-	-
Negative	63	-	64	2	-	57	217	-	181	2	-	47	36	-	1194
Adult	0	-	-	61	-	-	105	-	-	61	-	-	69	-	-
Illegal Device	1236	-	1235	1236	-	1242	977	-	1118	1236	-	**1252**	56	-	**1243**

When we look at the multiclass model (M), we can see from the Table 15 that the model has lower precision for the classes, but it also has a tendency to classify products

from adult class as Negative. The multiclass models (TD, M and E) for Illegal Devices, for example, had a misclassification greater than the binary, but errors were also thrown for the adult class, which would also be non-compliant (Figure 26). On the other hand, it has a tendency to classify Adult class products as Negative.

In Figure 25 it is possible to see items from the Negative class misclassified as Illegal Devices. They are, in this order: a smart video concierge, a car charger and an external router. Maybe, the shape of these images confused the classification models based on visual features.

Figure 25 – Negative product instances that were misclassified as Illegal Devices.

The same visual features could lead to the misclassification of the Illegal Devices in Figure 26, since the shape of these products do not give any clue about their illegality.

Figure 26 – Illegal Devices that were misclassified as Negative (the first one) or Adult (the last two).

We investigated the best models (see subsection 6.1) for `bin-adult` and `bin-illegal` in the divergence of classifications for the same products. We noticed that `TD bin-adult` exclusively hit only 2 products[1] in which `M bin-adult` and `E bin-adult` classified incorrectly.[2] When performing the opposite comparison[3] among the models we did not find any product.

[1] These products are not shown in this document due to be sensitivity of them, since they are true Adult products.

[2] The query for this is: `label==1 and adulttextual==1 and (adultmulti==0 or adultensemble==0`.

[3] The query for this is: `label==1 and adulttextual==0 and (adultmulti==1 or adultensemble==1`.

Chapter 7

Conclusions

This work discussed unimodal and multimodal approaches for detecting non-compliant products in a multiclass scenario: Negative, Adult and Illegal Devices. From the experiments presented here, the best models were the binary textual ones produced taking into account product titles and descriptions: `TD bin-adult` (with a recall of 98%, see Table 13) and `TD bin-illegal` (with a recall of 95%, see Table 13). This decision was made considering the good performance of these models regarding recall values but also the speed of prediction of only 0.5 seconds for almost 4,000 products.

Although TD (bin-adult and bin-illegal) models were chosen as the best ones, the multimodal (M) and ensemble (E) models had quite the same results for the classification task what could indicate that there is a small contribution of the visual modality to the classification task of non-compliant products in the experimented classes. This indicates that we should look for alternatives to improve the visual modality to re-evaluate multimodal approaches (M and E) in other scenarios, as well in other categories of non-compliant products.

7.1 Main contributions

This work proposed and investigated unimodal and multimodal approaches for the task of classification (or detecting) non-compliant products from Adult and Illegal Devices classes in a binary and also a multiclass scenario.

Thus, the main contributions of this work are:

- ❑ A comprehensive comparison between the different approaches (unimodal versus multimodal versus ensemble and binary versus multiclass).

- ❏ An error analysis tool that can be used to qualitatively investigate and compare model classifications in other works.
- ❏ The definition of non-compliant classification models, based on textual information, to be applied by the partner company.

7.2 Future work

As discussed in Section 6.1, the greatest contribution for the performance of the emsemble and multimodal models is given by the textual ones, but there is evidence that an improvement in the visual model could bring improvements for the M and E models. Therefore, studying new approaches to the visual model such as co-attention or training the model for the specific domain of e-commerce (with images of products), as have been done by related works, can contribute to an improvement in this modality and ultimately to multimodal learning.

There is also the possibility of testing these models in different categories of non-compliant products. According to the dataset of Americanas S.A., there are 19 classes of non-compliant products and this work only investigated two of them.

Another possible future experiment is to train a binary classifier for the Negative class (0) against the joining of the Adult and Illegal Devices classes as just one Positive class (1).

Finally, an improvement in the error analysis tool, with new functionalities to get insights of automatic validation of the models in visual and textual modalities could also be investigated.

References

AMOUALIAN, H. et al. An e-commerce dataset in french for multi-modal product categorization and cross-modal retrieval. In: HIEMSTRA, D. et al. (Ed.). **Advances in Information Retrieval**. Cham: Springer International Publishing, 2021. p. 18–31. ISBN 978-3-030-72113-8.

BALTRUsAITIS, T.; AHUJA, C.; MORENCY, L.-P. Multimodal machine learning: A survey and taxonomy. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 41, n. 2, p. 423–443, 2019.

BI, Y.; WANG, S.; FAN, Z. A multimodal late fusion model for e-commerce product classification. **CoRR**, abs/2008.06179, 2020. Disponível em: <https://arxiv.org/abs/2008.06179>.

BOJANOWSKI, P. et al. Enriching word vectors with subword information. **Transactions of the association for computational linguistics**, MIT Press, v. 5, p. 135–146, 2017.

CHEN, Y. et al. Destruction and construction learning for fine-grained image recognition. In: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2019.

CHOLLET, F. et al. **Keras**. 2015. <https://keras.io>.

CHORDIA, V.; KUMAR, V. Large scale multimodal classification using an ensemble of transformer models and co-attention. **CoRR**, abs/2011.11735, 2020. Disponível em: <https://arxiv.org/abs/2011.11735>.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE Conference on Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2009. p. 248–255.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

FORMAL, T. et al. Learning to rank images with cross-modal graph convolutions. In: JOSE, J. M. et al. (Ed.). **Advances in Information Retrieval**. Cham: Springer International Publishing, 2020. p. 589–604. ISBN 978-3-030-45439-5.

GAUTAM, H. **Word Embedding: Basics**. 2020. <https://medium.com/@hari4om/word-embedding-d816f643140>. Accessed: 2022-04-09.

GIRSHICK, R. et al. Rich feature hierarchies for accurate object detection and semantic segmentation. In: **Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2014.

HE, K. et al. Deep residual learning for image recognition. **CoRR**, abs/1512.03385, 2015. Disponível em: <http://arxiv.org/abs/1512.03385>.

_____. Deep residual learning for image recognition. In: **2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2016. p. 770–778.

KIM, Y. Convolutional neural networks for sentence classification. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1746–1751. Disponível em: <https://www.aclweb.org/anthology/D14-1181>.

KIROS, R.; SALAKHUTDINOV, R.; ZEMEL, R. S. Unifying visual-semantic embeddings with multimodal neural language models. **CoRR**, abs/1411.2539, 2014. Disponível em: <http://arxiv.org/abs/1411.2539>.

KRIZHEVSKY, A. Learning multiple layers of features from tiny images. **University of Toronto**, 05 2012.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. Imagenet classification with deep convolutional neural networks. p. 1097–1105, 2012.

LECUN, Y.; CORTES, C. MNIST handwritten digit database. 2010. Disponível em: <http://yann.lecun.com/exdb/mnist/>.

LIU, Y. et al. Roberta: A robustly optimized BERT pretraining approach. **CoRR**, abs/1907.11692, 2019. Disponível em: <http://arxiv.org/abs/1907.11692>.

LOCHTER, J. V. **Teoria de aprendizagem da leitura aplicada na inferência de palavras desconhecidas**. Tese (Doutorado), 2021.

LU, J. et al. Hierarchical question-image co-attention for visual question answering. In: **Proceedings of the 30th International Conference on Neural Information Processing Systems**. Red Hook, NY, USA: Curran Associates Inc., 2016. (NIPS'16), p. 289–297. ISBN 9781510838819.

MANIT, J.; SCHWEIKARD, A.; ERNST, F. Deep convolutional neural network approach for forehead tissue thickness estimation. **Current Directions in Biomedical Engineering**, v. 3, n. 2, p. 103–107, 2017. Disponível em: <https://doi.org/10.1515/cdbme-2017-0022>.

MCAULEY, J. et al. Image-based recommendations on styles and substitutes. In: **Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval**. New York, NY, USA: Association for Computing Machinery, 2015. (SIGIR '15), p. 43–52. ISBN 9781450336215. Disponível em: <https://doi.org/10.1145/2766462.2767755>.

MIKOLOV, T. et al. Distributed representations of words and phrases and their compositionality. In: **Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2**. Red Hook, NY, USA: Curran Associates Inc., 2013. (NIPS'13), p. 3111–3119.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://www.aclweb.org/anthology/D14-1162>.

ROMUALDO, A.; REAL, L.; CASELI, H. Measuring brazilian portuguese product titles similarity using embeddings. In: **Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana**. Porto Alegre, RS, Brasil: SBC, 2021. p. 121–132. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/stil/article/view/17791>.

SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. In: BENGIO, Y.; LECUN, Y. (Ed.). **3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings**. [s.n.], 2015. Disponível em: <http://arxiv.org/abs/1409.1556>.

SZEGEDY, C. et al. Inception-v4, inception-resnet and the impact of residual connections on learning. In: **Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence**. [S.l.]: AAAI Press, 2017. (AAAI'17), p. 4278–4284.

_____. Going deeper with convolutions. In: **2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)**. [S.l.: s.n.], 2015. p. 1–9.

WIROJWATANAKUL, P.; WANGPERAWONG, A. Multi-label product categorization using multi-modal fusion models. **CoRR**, abs/1907.00420, 2019. Disponível em: <http://arxiv.org/abs/1907.00420>.

WU, F. et al. Modality-specific and shared generative adversarial network for cross-modal retrieval. **Pattern Recognition**, Elsevier, p. 107335, 2020.

ZAHAVY, T. et al. Is a picture worth a thousand words? a deep multi-modal architecture for product classification in e-commerce. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 32, n. 1, Apr. 2018. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/11419>.

ZHANG, C. et al. Multimodal intelligence: Representation learning, information fusion, and applications. **IEEE Journal of Selected Topics in Signal Processing**, IEEE, 2020.

ZHOU, V. **A Simple Explanation of the Bag-of-Words Model**. 2019. <https://victorzhou.com/blog/bag-of-words/>.